Downloaded from orbit.dtu.dk on: May 05, 2023

DTU Library

Systems Biology of Metabolic Networks: Uncovering Regulatory and stoichiometric Principles

Patil, Kiran Raosaheb

Publication date:
2007

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Patil, K. R. (2007). *Systems Biology of Metabolic Networks: Uncovering Regulatory and stoichiometric Principles*.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Systems Biology of Metabolic Networks: Uncovering Regulatory and Stoichiometric Principles

Kiran Raosaheb Patil

,

Ph.D. thesis

Center for Microbial Biotechnology, BioCentrum-DTU,

Technical University of Denmark

Copyright ©2003-2006 Kiran Raosaheb Patil. All rights reserved.

Preface

Networks are natural aids to our understanding and depiction of all aspects of life. Biology is not an exception. Amongst several of the known biological networks, metabolic networks are probably the best studied networks in terms of physical laws governing their structure and operation. Furthermore, wide conservancy from microscopic bacteria to humans is what makes metabolic networks still more appealing to study in detail. Metabolic networks are comprised of the molecular machinery that processes food in order to generate energy and basic building blocks for growth. Consequently, cellular metabolic networks directly influence all other physiological processes, like it is said in India: "*we are what we eat*" and "*food is the most sacred*". Although much is known about different reactions that occur during metabolism, it is relatively poorly understood how these reactions are systematically regulated. Indeed, one of the most important and fascinating aspects of the metabolism is its adaptability and robustness towards different environmental and genetic changes. This flexibility of metabolic networks is due to several regulatory mechanisms. How this regulation is exerted over a complex network of reactions? Are there any simple and general rules governing their operation? How did such regulatory circuits emerge during evolution? This thesis is an effort towards answering these questions, and a result of little less than three years of my work as a PhD student.

Although human metabolic network would have been an obvious choice for this study, I have focused almost entirely on yeast metabolism. This is primarily because of the better knowledge-base available for investigating yeast metabolism and its regulation. Some of the general principles understood from yeast will form a basis for understanding operation of human metabolic networks. Nonetheless, this thesis makes a decent contribution towards gaining new insights into metabolic regulation and for exploiting this knowledge for designing microbial cells *in silico*.

The title of my PhD project in the beginning was "*Analysis of pathway structures in yeast*"; which then over time evolved to "*Systems Biology of Metabolic Networks: Uncovering Regulatory and Stoichiometric Principles*". Although this change also reflects the trend in current research, the driving force behind this change is not the *buzz words*, but rather the changes in my way of thinking about biology in general.

Although I wish and hope that this thesis will be a useful reading for people from many different backgrounds, some familiarity with basic biology jargon and understanding of some fundamental statistical concepts is perhaps taken for granted. I sincerely apologize if it is so, and in such case I wish that at least the general approach is clear enough to follow the work. Since my skills with

writing have not yet been tested for people outside a narrow field, it will be very nice to have your comments/suggestions on this thesis.

The title page of this thesis quotes a sentence from Lewis Carroll's *Alice's Adventures in Wonderland*. I read this marvelous book in the beginning of my PhD and it is one of my favorites. I have used several quotes from this book and from the other work of Lewis Carroll *Through the Looking Glass* throughout this thesis (in the beginning of most of the chapters and few other places in the beginning of the thesis). All quotes are thus borrowed from one of these two books. Several chapters also start with a picture/sketch. All photographs are taken by me during the period of this work, and none for this specific purpose. Although the pictures and the quotes are not directly related to each other, they are connected through the theme of the chapter. These pictures and quotes are complementing and sometimes partly concluding the story. Any way, *'what is the use of a book without pictures or conversations?'*

Acknowledgements

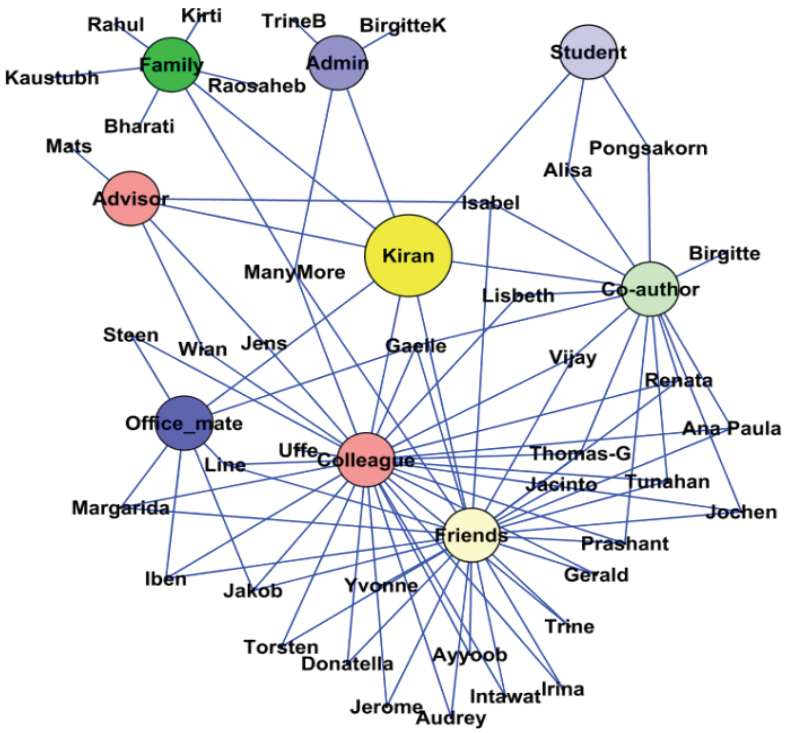
There are no exact words to express certain feelings. Unfortunately it is so for expressing the gratitude. This job is often done better by actions than words, and I hope that I have done that throughout. Nonetheless I take this opportunity to express my sincere thanks to all the people who have directly or indirectly helped and encouraged me during my PhD.

First of all, I would like to dedicate this thesis to three wonderful people, Jacinto, Ana Paula and Yvonne, without whom this work would have neither started nor finished.

I most cordially thank all of my family for being supportive, encouraging and always being positive about me.

Many thanks to Jens Nielsen who has not only been my advisor but also one of the most optimistic and encouraging persons I have met. He strongly supported my ideas and led me to the right track when I was lost.

There are so many other people that I wish to thank, I was afraid that it will need too much space to write it down here. These people include my colleagues, collaborators, friends, administrative and technical staff and so on. I finally decided to make an "acknowledgment network" (please see the next page). I am very grateful to all these people. They have all been very helpful and supporting. Special words of thanks to Prashant, Iben, Torsten and Trine for making my working (and non-working) time a pleasure.


Rahul
Kirti
TrineB
BirgitteK
Family
Admin
Student
Kaustubh
Raosaheb
Bharati
Mats
Pongsakorn
Alisa
Advisor
Isabel
Kiran
Birgitte
ManyMore
Co-author
Lisbeth
Steen
Wian
Jens
Gaelle
Vijay
Renata
Office_mate
Ana Paula
Uffe
Colleague
Thomas-G
Line
Jacinto
Tunahan
Margarida
Jochen
Friends
Prashant
Iben
Jakob
Yvonne
Gerald
Trine
Torsten
Donatella
Ayyoob
Irina
Intawat
Jerome
Audrey

Alice laughed. "There's no use trying," she said: "one can't believe impossible things." "I daresay you haven't had much practice," said the Queen. "When I was your age, I always did it for half-an-hour a day. Why, sometimes I've believed as many as six impossible things before breakfast."

Synopsis

Understanding the general principles governing the functioning of metabolic networks is a major objective of the work presented in this thesis. Functionality of metabolic networks can be viewed from two aspects, *viz.*, stoichiometric and regulatory. All possible modes of operations of metabolic networks (feasible flux space) are confined by the stoichiometry (and thermodynamics). Regulation then imposes additional constraints that determine which of the numerous possible phenotypes is observed under a given condition. It is relatively easy and intuitive to understand and interpret stoichiometric aspects of metabolic function in light of mass and (to limited extent) energy balance laws. In contrast, operation of regulatory circuits and their effects on cellular operations has been to large extent a descriptive science. Nonetheless, new opportunities to uncover and test principles of regulation are being opened through the availability of large amount of molecular abundance data on genome-wide scale. Several of the existing approaches in this direction, however, are data-driven and thus lack potential to be generalized and extrapolated to different species. Thus new algorithms built on hypotheses and data rather than data-only are necessary for integration of omics data and discovery of biologically meaningful patterns.

Cellular response to genetic and environmental perturbations is often reflected and/or mediated through changes in the metabolism. Such metabolic changes are often exerted through transcriptional changes induced by complex regulatory mechanisms coordinating the activity of different metabolic pathways. It is difficult to map such global transcriptional responses by using traditional methods, because many genes in the metabolic network have relatively small changes at their transcription level. I therefore developed an algorithm that is based on hypothesis-driven data analysis to uncover the transcriptional regulatory architecture of metabolic networks. By using information on the metabolic network topology from genome-scale metabolic reconstruction, it is shown possible to reveal patterns in the metabolic network that follow a common transcriptional response. Thus, the algorithm enables identification of so-called reporter metabolites (metabolites around which the most significant and coordinated transcriptional changes occur) and a set of connected genes with significant and coordinated response to genetic or environmental perturbations. The results imply that cells respond to perturbations by changing the expression pattern of several genes involved in the specific part(s) of the metabolism in which a perturbation is introduced. These changes then are propagated through the metabolic network because of the highly connected nature of metabolism. Although structure of the regulatory network determines the details of how the transcriptional regulatory program is executed, the meta-

bolic network itself seems to guide this machinery, which we see as the consequence of the fact that metabolic regulation has been designed and evolved *for and around* the metabolites.

Since flow of mass through each metabolite is subjected to stoichiometric and thermodynamic constraints, I argue that the coordinated expression of the genes surrounding a metabolite is partly a thermodynamic necessity. Is this requirement reflected in the evolution of metabolic genes? This question was addressed through systemic analysis of gene modules emerging from the metabolic network topology with respect to their sequence evolution rates, shared promoter sequence motifs and transcriptional co-regulation. The sequence conservancy was found to be significantly over-represented in gene modules associated with the metabolites that are crucial for survival of yeast. It is further shown that several of these gene modules share sequence motifs in their promoter regions and also exhibit a high degree of transcriptional co-regulation evaluated across a large gene-expression dataset. These results imply that the topology of metabolic network constraints the evolution of regulatory circuits. In yeast some of these regulatory circuits are built around the evolutionary conserved metabolic neighbors. Thus network topology sheds light on the link that connects organization and operation of regulatory circuits to the evolution of DNA sequences. This is a significant step forward in terms of understanding the emergence of regulatory circuits, whose existence is generally taken for granted. Some of these regulatory circuits are closely knitted over and constrained by the network topology, which signifies mass and energy balance constraints.

Another outcome of the work presented in this thesis is the demonstration of the importance of highly connected metabolites in regulation and functionality of the metabolic network. Many of the highly connected metabolites (such as redox and energy co-factors) are usually omitted *a priori* from the analysis of transcriptome and other omics data. Here it is shown that not only such omission is unnecessary, but it may also critically affect the results obtained. Since highly connected metabolites glue the network together, their role in terms of transcriptional regulation is significant for understanding and interpreting global changes in the network.

The algorithmic platform developed for integration of transcriptome data was extended to metabolome data. This analysis enabled the deduction of whether the regulatory response at the level of each reaction was predominantly subjected to hierarchical or metabolic regulation. Finally, the thesis also presents a new algorithm (OptGene) for exploiting stoichiometric operational principles of metabolic networks for *in silico* identification of metabolic engineering targets. Computational efficiency and flexibility of the OptGene algorithm makes it a versatile and useful tool for identification and screening of large number of metabolic engineering strategies. Indeed,

some of the strategies identified *in silico* for improved succinic acid production in yeast were successfully verified *in vivo*.

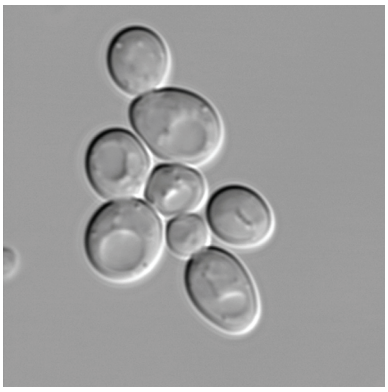
Overall, the work presented in this thesis comprises significant conceptual and algorithmic advances towards uncovering operational and evolutionary principles underlying complex metabolic networks.

Dansk sammenfatning

I forbindelse med industriel udnyttelse af mikroorganismer til produktion af kemikalier er der stor interesse for at analysere funktionen af metabolske netværk, herunder specielt opnå indsigt i hvordan aktiviteten af forskellige grene i sådanne netværk er reguleret. Idet metabolismen spiller en central rolle i funktionen af alle levende celler, er der også stor interesse for metoder til analyse af metabolske netværk i naturvidenskabelige studier samt i forbindelse med kortlægning af mekanismerne bag stofskifte relaterede sygdomme. I dette studium er der udviklet forskellige nye bioinformatiske metoder til analyse af levende cellers stofskifte. Metoderne er baseret på anvendelsen af metabolske modeller til integration af både transkriptionsdata og metabolome-data. Således er det vist at cellers respons til ændringer i deres miljø eller deres genetiske baggrund, i stor udstrækning involverer koordineret ekspression af et stort antal gener der er involveret i specielle metabolske moduler. Dette peger på at metabolismen er involveret i overordnet regulering af cellers respons til ændringer i deres miljø, en hypotese der er underbygget af at gener der koder for centrale enzymer i cellens metabolisme er evolutionsmæssigt mere konserveret end gennemsnittet af gener i genomet. De udviklede metoder er illustreret anvendt på en række data, specielt fra studier af bagegær men også for andre organismer, og har i flere tilfælde ført til betydelig ny biologisk indsigt i hvordan metabolismen er reguleret. I et videre arbejde er det demonstreret hvordan metabolske modeller kan anvendes til design af nye cellefabrikker. Dette har involveret udvikling af en ny computer algoritme der identificerer hvilke gener der skal modificeres for at opnå en forbedret cellefabrik. Anvendelse af denne algoritme demonstreret anvendt for flere eksempler, og videre eksperimentel verifikation en computer forudsagt design strategi er kort beskrevet. Afhandlingen demonstrerer derfor anvendelse af metabolske modeller både til analyse af store datasæt, til anvendelse i systembiologi, og til design af nye cellefabrikker, til anvendelse i metabolic engineering.

Contents

Chapter 1: Introduction



The White Rabbit put on his spectacles. `Where shall I begin, please your Majesty?' he asked. "Begin at the beginning,", the King said, very gravely, "and go on till you come to the end: then stop".

1.1 "Systems" and "Systems Biology"

The use of a well defined *system* (and thus *system* boundaries) for analyzing and predicting behavior of physical *systems*[1] was a standard practice taught to me in various thermodynamic, physics, mathematics and chemical engineering courses. However, it took a while before I realized that defining a system is not only one of the ways to analyze a scientific question but is the only way to do so. The origin of this fact does not lie in the nature of a particular problem, but in the very nature and limitations of human beings in terms of abstracting the reality for creating a convenient perception of the *universe*[2]. Consequently, any problem of either describing or predicting a phenomenon inherently implies the definition of a particular set of objects and boundaries, and thus, a system under investigation. The question of whether a system under investigation is always clearly and completely defined is both subjective and dependent on the nature of the problem. In fact, such difficulties and ambiguities force certain assumptions in order to simplify the problem conceptually and help defining a system, at least to some extent. Following are certain definitions of a system as documented at Wikipedia[3]:

System: System (from the Latin (systēma), and this from the Greek σύστημα (sustēma)) is an assembly of elements comprising a whole with each element related to other elements. Any element which has no relationship with any other element of the system, cannot be a part of that system. A subsystem is then a set of elements which is a proper subset of the whole system.

System (Thermodynamic): A thermodynamic system is defined as that part of the universe that is under consideration. A real or imaginary boundary separates the system from the rest of the universe, which is referred to as the environment or surroundings (sometimes called a reservoir.) A useful classification of thermodynamic systems is based on the nature of the boundary and the quantities flowing through it, such as matter, energy, work, heat, and entropy. A system can be

[1] Recursive nature of this statement is discussed in the following text.
[2] Here I take freedom to leave the task of defining boundaries for "universe system" up to the readers.
[3] http://en.wikipedia.org/wiki/Main_Page

anything, for example a piston, a solution in a test tube, a living organism, or a planet, etc. See figure 1.1 for an example of a thermodynamic system.

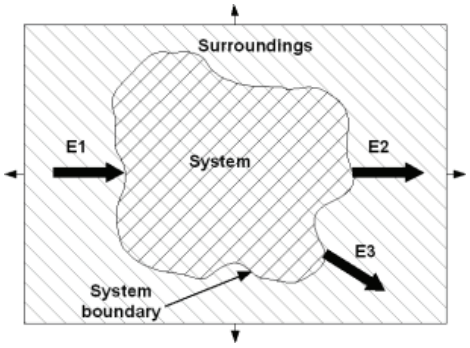


Figure 1.1. General concept of a system as used in thermodynamics. If E1, E2 and E3 denote the energies crossing the system boundary, then the energy balance equation for this system can be written as E1 = E2 + E3.

An interesting part of these definitions is the underlined text "a system can be anything". This statement clearly reflects the difficulties encountered in understanding the concept of a system. However, if we accept the hypothesis discussed in the beginning regarding nature and limitations of our perception about the universe, this apparent ambiguity can be cleared. Thus any definition of the systems is ambiguous by nature, but the extent of ambiguity decreases with imposed (either forced or conveniently defined) assumptions and restrictions on the desired accuracy of the descriptions/predictions.

The most striking difference between the above two definitions of systems is the *a priori* condition of relatedness forced by the first definition. This plainly contradicts "a system can be anything". I will adapt the thermodynamic definition of the system in light of my views regarding the nature of systems. Moreover the first definition is very imprecise as:

1) It does not define "element".

2) It does not define what kind of relations is expected between the elements.

3) Even though nature of relationship between elements is defined, it does not explain whether such relationships can/should always be established *a priori*?

4) Most importantly, it does not force existence of any boundaries, and thus makes it conceptually difficult to use in a systematic way.

However, the second definition is the one that may seem closer to the one **used** in this thesis. This is because although the first definition fails to define a system, it helps to **represent** a system, **once defined**.

Examples of the usefulness of a thermodynamic definition of a system are widespread in several engineering disciplines (classical example is the Carnot heat engine to produce work by using thermal energy). In contrast, it is only recently that the term systems biology is being used extensively[4]. Since the title of this thesis starts with the term "Systems Biology", I will try to justify its use here by attempting to define this term from my viewpoint. First I maintain that "Systems Biology" stands for "study of biological entities as systems". Naturally, the recursive nature of this definition (here due to the term *entity*) can not be avoided as for the general concept of a system. In this spirit it can be argued that all biological study is systems biology (and similarly all sciences are systems sciences) since one can not even define a scientific problem without having a definition of the underlying system (although this is not always done explicitly). Although such an argument is completely valid from a definition point of view, it is possible to draw a boundary (at least to a certain extent) between biology and systems biology based on the application (and mode of application) of the assumptions denominating the systems under investigation. The first definition of a system can help to draw such boundary as it gives a convenient representation of a well defined system.

Any investigation of a biological system that addresses and accounts for the defined interactions between defined components under investigation (in terms of their contribution in describing/predicting the behavior/properties of the system) can be called a systems biology problem.

The word "defined" should be emphasized here, as it indirectly sets the boundaries of a system and thus makes the definition more complete. The need for this restriction may not be apparent in many cases. However, if we think about all potential unknown factors that might affect the properties and behavior of the systems (e.g. unknown/weak interactions), the need for clearly

[4] Google search for "Systems Biology" yielded 6.66 million results on 17 July 2006.

defining the boundary of a system is immediately apparent. This definition is rigorous and other generally accepted definitions of systems biology can be seen as special instances of this definition. The definition presented here is not intended to classify existing biological research, but only to justify and clarify usage of this term in the title this thesis. This is important as, e.g., according to this definition mere measurements of certain biological properties for large number of molecules/cells (e.g. measurement of growth rates for many mutants, measurement of mRNA/protein/metabolite abundances on a genome-scale) does not fit into here-defined definition of systems biology. However, if the same experiment and its outcome are analyzed and defined from hypothesis driven systems perspective, e.g. as quantitative description of the behavior of hypothesized connectivity between components (e.g. genes); it will then fall under here-defined definition of systems biology.

1.1.1 Implications of systemic analysis for solving biological problems

The rigorous definition of a system is difficult to use in biological problems due to the inherent complexity of the components and interactions. It may also be argued whether it is really necessary, as the system level analysis is always inherently accounted for, whether rigorously defined or not. The real implications of using rigorous definitions can be exemplified by analogy with other fields of science where it has proven to be extremely useful. Such disciplines include electrical engineering, mechanical engineering and computer engineering. I will use an example from software programming to partially illustrate the applicability of a systems approach. One of the most popular programming paradigms used in modern software engineering is object-oriented programming. The notion of object-oriented programming is centered on the concept of an object. An object is a conceptual part of a program that performs a specific defined task. It differs from the traditional function in the sense that the object can have its own data and functions that can be made invisible for the rest of the program. Moreover, such objects can be used to create more objects with inherited properties. This simple concept brought a revolution in programming practice in terms of improved program clarity and reusability of code. Thus, a new program can be created by fitting together different old objects; like lego bricks fitted together to build toys. Similar concepts are also used in electrical engineering, e.g. different standard devices such as transistors, resistors, condensers etc. are used to build many different devices such as televisions, radios and computers. In the field of mechanical/construction engineering certain standard designs of small parts are used in many devices and automobiles. Furthermore, in chemical engineering standard unit operation devices such as distillation columns, filters etc. are used in almost all chemical plants. Although design details of a part/object/unit can vary depending on the need, generalized functional assignment to objects offers enormous help in designing complicated sys-

tems in short time and allows the use of previous experience to be used efficiently in designing new products/processes (e.g. imagine the time that will take to design a new car from scratch without using standardized parts).

If examined closely, it can be seen that this advantage of object-oriented design can be attributed to the efforts made to carefully define the functionality of each part in terms of their input/output parameters (system boundaries) and their general design (system architecture). I thereby argue that similar efforts in systems biology will help us in the future to put together different levels of knowledge to get a holistic picture of the cellular machinery. An ancient Indian story of an elephant and blind men (Figure 1.2) is a good illustration of the need for systemic integration. According to the story, six blind men were trying to describe the shape of an elephant by touching it. However, since each man was feeling different part of the elephant's body, each of them described it as a different shape. A wise man helped them to understand that all of them were right and the elephant has all those shapes. Thus, the story implies that the complete picture of the reality can be obtained only by integrating knowledge from different levels of analysis. In consequence and parallel, this process of systemic integration will also help us to design biological parts and whole systems with desired functionality. A scientific discipline referred to as synthetic biology (Bork, 2005)[5] aims at achieving this goal.

[5] Present day synthetic biology is still in the very beginning of this process.

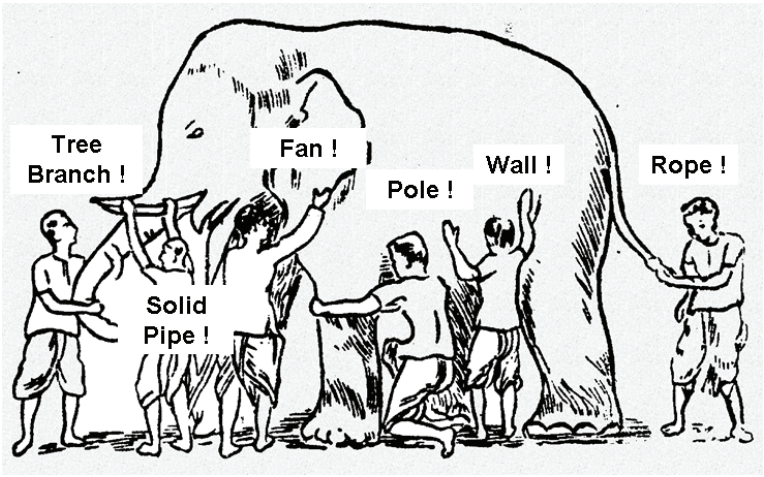


Figure 1.2. Depiction of an ancient Indian story where six blind men are trying to understand the structure of an elephant by feeling different parts of an elephant. The story conveys the idea that the true picture of reality can only be perceived through the integration of all different view points.

1.2 Miscellaneous background information

1.2.1 Molecular operations of cell and metabolism

Modern cellular biology is usually understood and studied in light of a flow diagram (based on central dogma (Crick, 1970)) depicting how information encoded in genes (genotype) is reflected at the level of cellular function and state (phenotype) (Figure 1.3a). Information stored in the genes is in the form of sequence order comprised of four nucleotides (**A**denine, **T**hymine, **G**uanine and **C**ytosine). A triplet (e.g. ATG) codes for one of the 20 amino-acids. Several of the amino acids joined together constitute a protein. Thus a gene codes for a specific protein via transcription (DNA to mRNA) and translation (mRNA to protein).

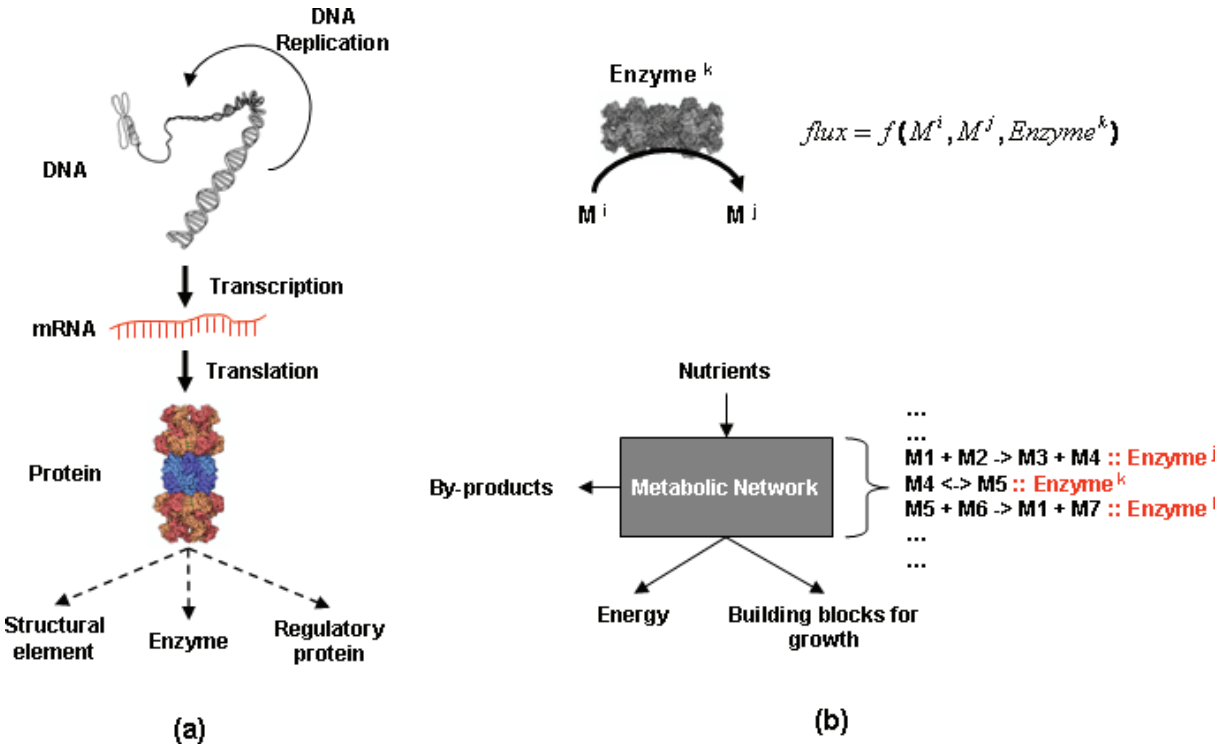


Figure 1.3. (a) Central dogma in molecular biology. DNA replication can be thought as information flow (back-up) from genome to genome. Information coded in genes flows to proteins via transcription and translation. Proteins may play variety of functional roles in a cell. Only three roles are shown as examples. (b) Enzymes catalyze chemical transformation of metabolites. The rate of enzyme catalyzed reaction (flux) is not only a function of enzyme availability and properties, but also concentration of substrates and products. Several of such enzymatic steps constitute a metabolic network where products of some reaction (/s) serve as substrates for other reaction (/s), thus creating an interconnected reaction web. The overall function of a metabolic network can then be viewed as utilizing environmentally available nutrients to generate energy and building molecules for growth and maintenance of the cell.

One of the important roles[6] played by proteins is as catalyst for chemical and/or physical transformation of various chemical substances. Such proteins are usually referred to as enzymes. Together, these enzymes create a network of reactions where substrates (food) available in the environment is broken down to generate energy and building block molecules. Generated energy is

[6] Other example roles for proteins include signal transduction, transcriptional regulation, structural element etc.

then used to assemble these building blocks towards creating new cells and for maintaining the existing cells. The whole process is termed metabolism which operates through a metabolic network (Figure 1.3b). The term metabolite is generally used to refer to only relatively "low" molecular weight compounds and excludes all cellular substances that are genetically encoded (e.g. RNA and proteins) (Jewett et al., 2006).

1.2.2 Structure and operation of metabolic networks

Remarkably, the basic architecture of metabolic networks is largely conserved across several different species ranging from microscopic bacteria to plants and humans (Peregrin-Alvarez et al., 2003). Thus the cellular machinery fueling distinct functionality and phenotypes is founded on identical metabolic processes. Understanding of these general organizational principles of metabolic networks can be facilitated by graph-theoretical representation of metabolic networks (Figure 1.4). Enzymes and metabolites can be viewed as nodes in this network, while interactions between them form edges. The number of neighbors for a node is referred to as its degree. The study of network properties based on its connectivity is called topological analysis. Metabolic networks generally form a fully connected network, i.e., it is possible to travel from any node to all other nodes in the network. The number of edges traversed in such path is referred to as the distance between two nodes. Metabolic networks in different species show similar scale-free topology (Strogatz, 2001) where few metabolites are involved in large number of reactions (also known as hubs), while most of metabolites take part in small number of reactions (Jeong et al., 2000). Metabolic hubs also bestow interesting small-world property to these networks (Fell and Wagner, 2000); meaning no two nodes (enzymes or metabolites) in a metabolic network are too far from each other. For example, any two nodes enzymes/metabolites in yeast metabolic network are on average 5 edges away from each other. Nevertheless, the high connectivity in metabolic networks must always be seen only in the context of stoichiometric restrictions on the flow of materials. In this aspect, metabolic networks differ from other networks such as protein-protein interaction networks, electrical grids and internet.

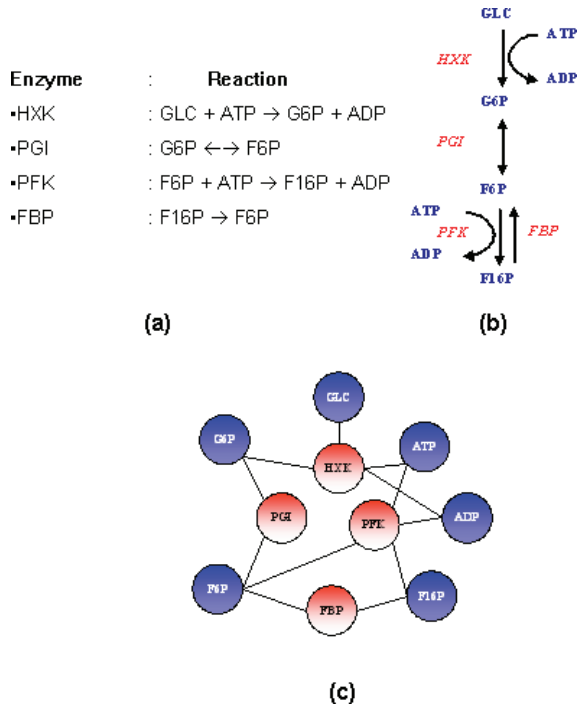


Figure 1.4. Graph theoretical representation of metabolic networks. A group of reactions (a) are generally represented as pathways (b) where co-factors and other highly connected metabolites are depicted only at individual reaction level. (c) Graphical representation where enzymes/reactions and metabolites form nodes and interaction among them as edges. Such graph is essentially bi-partite, as neither enzyme nor metabolite nodes are directly connected to other nodes in the same category.

Metabolic networks across different species are, however, different in many aspects. Differences are more evident at the level sequences/structures of individual enzymes and the regulation of enzymes in response to environmental/genetic challenges. Regulation of enzyme production is necessary for an organism in order to: i) allocate resources optimally so as to produce only enzymes that are necessary under given conditions and only in required amounts, and ii) avoid excess (or too less) amounts of enzymes which may result in unbalanced distribution of substrates that are entering the cell. In order to further elaborate on the concept of metabolic regulation, it is necessary to define the term "flux". In metabolic context, flux for a certain reaction refers to

the amount of substrates processed (or products produced) per unit time[7]. The integrated effect of fluxes through different reactions in a network can be observed in the phenotype of a cell (e.g. amounts of substrate consumed and final products and by-products formed, growth rate of an organism etc.). Due to the highly interconnected nature of metabolic networks and stoichiometric constraints, fluxes are to a large extent inter-dependent. Indeed, inter-dependencies between fluxes can be systematically mapped in a flux-coupling graph (Burgard et al., 2004) where two flux nodes influencing each other are connected by an edge. Interestingly, although not unexpectedly, the flux-coupling graph also shows a scale-free topology marked by few hub fluxes (Burgard et al., 2004).

Organization and operation of metabolic networks is traditionally viewed and understood in terms of ensemble of pathways. Pathways are comprised of groups of enzymes acting towards production/breakdown of certain metabolites (/group of metabolites). Glycolysis and TCA cycle are familiar examples of pathways. Although the concept of a pathway is widely used and very useful for pictorial representation, the definition of a pathway is very vague from stoichiometric point of view. An alternative and more comprehensive way of understanding the operation of metabolism is through the enumeration of all possible combination of reactions that can sustain a balanced flow from substrates to final products (Schuster et al., 2000). Each such combination (termed Elementary Flux Mode) can then be seen as a definition of a stoichiometrically "complete" pathway. An important fact is that the number of such possible routes is of the order of millions, even in a simple microbial metabolic network (Klamt and Stelling, 2002). Any active metabolism at a steady state can be represented as a linear combination of these elementary flux modes. What are the factors/mechanisms responsible for particular phenotype under given conditions? It appears that this choice is achieved via coordinated regulation of enzymes. Thus, not only a single enzyme is regulated for its optimal operation (for example see (Dekel and Alon,

[7] It is customary and convenient to normalize this quantity with respect to a certain measurable parameter. For example, ethanol produced per unit time per unit biomass.

2005)), but also the whole network is subject to regulation for optimal network functionality (for example see (Ibarra et al., 2002)).

Regulation of enzymes can occur at the level of transcription, translation or post-translational modifications (e.g. phosphorylation). Furthermore, enzyme activity may be regulated by small effector molecules. The flux through a reaction is dependent not only on the availability and activity of the enzyme, but also on the concentrations of substrates, products and effectors (Nielsen and Oliver, 2005) (Nielsen, 2003). Such relationships are usually non-linear. Additionally, due to the interconnected nature of the metabolic network, all steps in the metabolism can in principle influence all other steps. Consequently, understanding, simulation and prediction of both dynamic and steady state operations of metabolic network are challenging tasks.

1.3 Statistics/probability concepts used

1.3.1 Student's t-test, p-value and Z-score

One of the common problems encountered in biological data analysis is the comparison of features measured between two conditions or strains. For example, it is often needed to decide whether the expression of a gene has altered between the reference and the modified strain. Since the expression of a gene can vary to a certain extent even within the same strain (or under identical experimental conditions), it is necessary to repeat the experiment several times to know the distribution of its expression level. Student's t-test can then be used to compare, for example, expression levels of a gene between two conditions. The t-test provides a p-value which signifies the confidence that can be placed on the hypothesis that the expression has changed significantly. A p-value of 0.05 means, with 95 % confidence, that the expression of the gene is changed. In other words, p-value denotes the probability that the observed difference in the mean expression levels under two conditions is simply by chance. Since the t-test is based on the assumption of normal distribution of variables, in the case where this assumption is not valid, other statistical tests (e.g. rank-sum test) should be used.

A p-value can also be interpreted as a fraction of the area under a standard normal distribution curve (the total area equals 1 and so does the probability of any event for which the p-value is calculated). Thus, by using a cumulative distribution function (CDF) for the standard normal function, p-value can be converted to a Z-score (figure 1.5). Note that the x-axis is (1-p), which results in higher Z-score for low p-values.

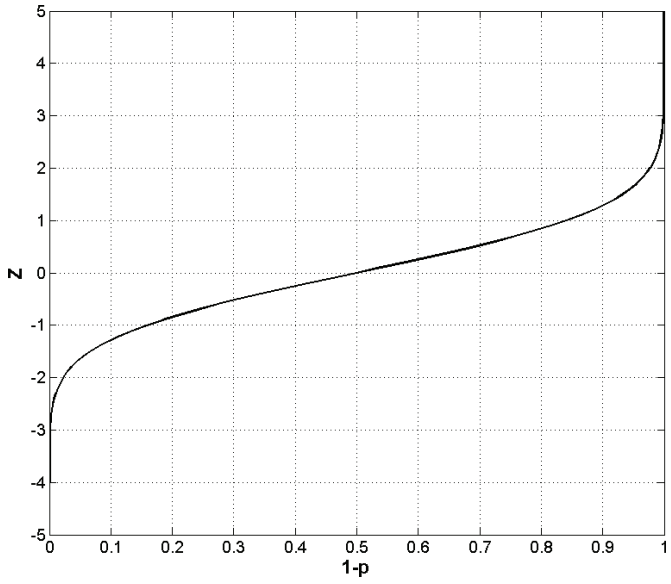


Figure 1.5. Normal cumulative distribution function (CDF).

1.3.2 Central limit theorem

Problems dealing with the large biological networks often require calculating the properties of a group of entities (e.g. genes) rather than single element. The simplest way to do this is to calculate the mean or average value of the property for the group under consideration. Thus, it is of interest to know how the distribution of mean values will look like for a sample of given size. Once this distribution is known, it can be used to access whether the mean value of the group under consideration is significantly higher (or lower) than the expected value. An example could be a question such as whether the fold changes in the expression levels of all genes in the TCA cycle (let's say 20 genes) is significantly higher than expected in a given experiment. If the distribution of the fold changes for individual genes (population) is normal, the resulting distribution of means of gene groups of size 10 (samples) will also be a normal and a p-value can be easily calculated by using normal CDF. In case where the distribution for the population is not normal, calculation of p-value is not trivial. However, the central limit theorem states that irrespective of the distribution of the population, the distribution of means always tends to be normal. The distribution approaches to normality with the increased group size. The mean of the samples equals to the population mean, while the variance equals [population variance/group size]. Thus, even when the data is not normally distributed, the p-value based on the distribution of means can be calculated easily by using the central limit theorem and normal CDF.

1.3.3 Pearson correlation coefficient

Pearson correlation coefficient indicates the strength and direction of the correlation between two variables. This coefficient is often used to measure the strength of correlation between expression levels of two genes. Significantly correlated genes may imply a common biological mechanism governing their expression or a close functional relationship.

1.4 Overview of the thesis

Metabolism is one of the key cellular processes providing necessary precursor molecules and free energy necessary for biosynthesis and maintenance. This central role of metabolism is evident by two facts. Firstly, several of the metabolic pathways are well conserved across different domains of life (Peregrin-Alvarez et al., 2003); and secondly, the cellular response to genetic and environmental perturbations is often reflected and/or mediated through changes in the metabolism. Consequently, it is not surprising that several diseases (e.g. diabetes, cancer, obesity) are closely associated with metabolic disorders/malfunctioning. Moreover, the metabolism of microorganisms is largely used as cell factories for producing a variety of chemical and pharmaceutical products. Recently, mammalian cell cultures are also being used for producing several products through cellular metabolism. Thus, understanding of organizational and functional principles of metabolic networks is an essential pre-requisite for devising rational strategies, not only for combating diseases, but also for metabolic engineering of the cell factories.

Owing to the topological and regulatory complexity of metabolism, emergent systemic properties play as important role in the operation of metabolism as the properties of its constituting components. My research ideas are hence centered on understanding the biological logic behind large-scale organization, operation and design of cellular metabolism from a systems perspective. Since cellular metabolism, as reflected in the metabolite levels and fluxes, is an integrated result of mass balance constraints and regulation, the organization of this thesis can also be broadly classified into stoichiometric (& topological) and regulatory analysis of metabolism (Table 1.1). Stoichiometry represents the mass balance constraints on a metabolic network at (pseudo-) steady state and can be viewed as limits on the possible operational modes in n-dimensional flux space. Regulatory networks impose additional constraints on this solution space and thus together these two decide what metabolic phenotype will be observed under given conditions (Figure 1.5). For the sake of simplicity and practicality, the analysis presented in this thesis is focused on the (pseudo-) steady state operation of metabolism. Such assumptions are to a large extent justifiable for several applications concerning fast growing microbial systems.

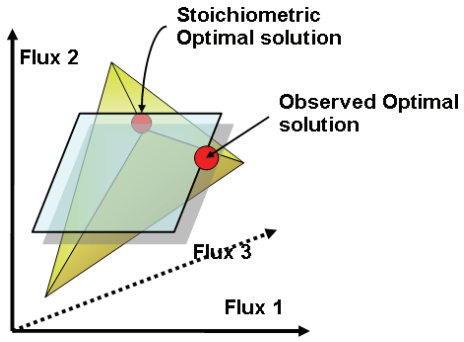


Figure 1.5. Constraints on the metabolic phenotype. This schematic shows a metabolic system consisting of three fluxes. Stoichiometric constraints (originating from mass balance law) define feasible space (yellow cone), while regulation imposes additional constraints (blue plane) and in result may alter the observed phenotype. The optimal solution refers to observed phenotype under the assumption that cellular metabolism operate so as to optimize a certain task (objective function).

A major part of this study is devoted to the analysis of transcriptional regulation in metabolic networks. A question that attracted my attention was how cells respond (or will respond) to a given perturbation if such information is not coded in the genome *a priori*. The number of such possible disturbances is practically infinite (Figure 1.6).

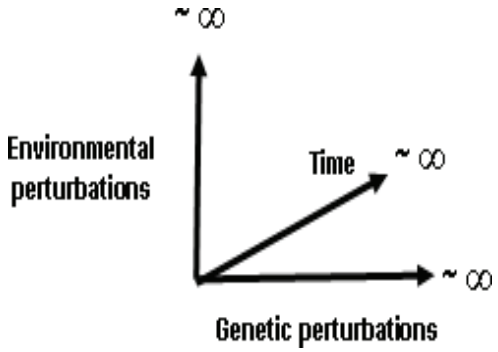


Figure 1.6. Perturbation space for biological systems. A cell can be subjected to any of, or combination of, perturbations at genetic or environmental level. The disturbance can also be a function of time. Since all axes have practically infinite limits, it is interesting to study how cells with finite memory and capabilities respond to perturbations in orderly manner.

It can be deduced by induction that the space of possible transcriptional responses will also be infinite, as the space of possible disturbances. Furthermore, for many of such unexpected disturbances the transcriptional response has been found to be altered rationally. Since there is no way

that all possible responses can be coded in a genome of a finite size, the way that infinite responses can be generated has to be based on the guiding principles or rules that are coded in the genome. In the case of transcriptional responses in metabolic networks, the metabolic network itself is a common denominator for all perturbations (and response to perturbations). Hence the principles of transcriptional regulation in a metabolic network must stem from network properties, e.g. the topology of the network. Thus most of the transcriptional analysis presented in this thesis is knitted over the topology of metabolic networks.

The living matter as we see it today is a result of (at least) several millions of years of evolution. This evolution is generally thought of as an optimization process. Objective functions of evolutionary optimization are not immediately evident, partly due to incomplete knowledge about present state and partly due to insufficient information about past environment and genetic factors influencing the evolution. However, it is clear that the optimization process is still underway and can be understood and manipulated in terms of its objective function in some simple systems (Dekel and Alon, 2005) (Ibarra et al., 2002) (Chatterjee and Yuan, 2006). On the other hand, it is thought that several biological systems have already reached a close-to-optimal solution. In either case, the properties, behavior and response of living matter can only be understood in light of its past. Much like a country or a culture can not be understood without understanding its history. Thus regulation of metabolism can only be completely understood in terms of the evolution of the metabolic network. Consequently, part of this thesis is also dedicated towards understanding the evolutionary basis underlying mechanisms and organization of metabolic regulation.

Table 1.1 puts different chapters in this thesis into three broad categories, *viz.*, stoichiometry centered analysis, regulation of metabolism and evolution of metabolism. Chapter 2 reviews the current research on the use of genome-scale models in metabolic engineering. Metabolic engineering is a science of (re-) designing metabolic networks so as to obtain a desired metabolic phenotype, e.g. increased production of metabolites from microbial cells. Metabolic engineering thus represents one of the problems where systems biology approach will be of great use. Genome-scale models of metabolism form a basis for such system level analysis of metabolic networks as a whole. Indeed, most if the subsequent work is based on the use of genome-scale metabolic models as a scaffold for systems analysis of metabolic networks. Most of the discussion in chapter 2 is stoichiometry related as the primary use of genome-scale models so far has been for stoichiometric analysis.

Chapter 3 introduces a novel algorithm (reporter algorithm) where genome-scale metabolic models have been brought to the forefront of the transcriptome data analysis. This chapter also pre-

sents a fundamental algorithmic structure for conceptual understanding of organization of transcriptional response in metabolic networks. Chapter 4 discusses some improvements over the part of the algorithm presented in the previous chapter. The next chapter (Chapter 5) extends the reporter algorithm for the analysis of the metabolome data. Furthermore, it also discusses a strategy for deducing the logic of multi-level regulation in metabolic networks by combined analysis of the transcriptome and the metabolome data.

The sixth chapter addresses the problem of metabolic regulation from the evolution point of view. This work also uses a large scale transcriptome data to deduce significant regulatory patterns. Two additional gene expression datasets are also analyzed to complement the hypothesis of metabolite centered regulatory architecture. Once strong evidences of metabolite centered regulation has been furnished in the previous chapters, chapter 7 then provides a more detailed analysis of metabolome data and draws certain general conclusions regarding the nature of regulation in metabolic networks.

Chapter 8 presents an algorithm (OptGene) that exploits stoichiometric operating principles of metabolic networks in order to identify metabolic engineering targets. Although no regulatory information is directly included in the analysis, the OptGene algorithm certainly forms a platform for devising an extensive methodology where stoichiometric analysis is married with the regulatory principles learned in the previous chapters. The next chapter briefly discusses the work on experimental verification of some of the strategies identified by the OptGene.

Chapter 10 summarizes a few other small stories built along with the main work in the thesis. Although, most of the work is still ongoing, it deserves a place in this thesis for the sake of completeness. Finally, chapter 11 runs through some of my thoughts about directions of future research in the field of metabolic systems biology.

Table 1.1 Broad classification of the work presented in this thesis.

Theme	Chapters
Stoichiometry centered analysis	1,2,8,9
Regulation of metabolism	1,2,3,4,5,7,10
Evolution of metabolism	1,6,10

Chapter 2: Use of genome-scale microbial models for metabolic engineering

This chapter is based on the publication:
Patil, K. R., Akesson, M. & Nielsen, J. Use of genome-scale microbial models for metabolic engineering. Current Opinion in Biotechnology 15, 64-69 (2004).

"Then you should say what you mean." the March Hare went on. "I do," Alice hastily replied; "at least -- at least I mean what I say -- that's the same thing, you know." "Not the same thing a bit!" said the Hatter, "Why, you might just as well say that 'I see what I eat' is the same thing as 'I eat what I see'!" "You might just as well say," added the March Hare, "that 'I like what I get' is the same thing as 'I get what I like'!" "You might just as well say," added the Dormouse, who seemed to be talking in his sleep, "that 'I breathe when I sleep' is the same thing as 'I sleep when I breathe'!" "It is the same thing with you," said the Hatter, and here the conversation dropped, and the party sat silent for a minute

2.1 Summary

Metabolic engineering serves as an integrated approach to design new cell factories by providing rational design procedures and valuable mathematical and experimental tools. Mathematical models play an important role for phenotypic analysis, but they may also be used for design of optimal metabolic network structures. The major challenge for metabolic engineering in the post-genomic era is to broaden its designing methodologies to incorporate the biological data on whole genome-scale, and genome-scale stoichiometric models of microorganisms represent a first step in this direction.

2.2 Introduction

The metabolic capabilities of different microorganisms for producing valuable compounds are widely exploited in the pharmaceutical and chemical industry. Furthermore, there is an increasing trend to replace chemical production processes with biotech routes based on microbial fermentations. As microorganisms are typically evolved for survival and growth in their natural habitat, it is often necessary to retrofit the genotype of the applied cell factory to obtain a desired phenotype. Through the use of directed genetic modifications using genetic engineering it has become possible to realize the metabolic potentials of many different microorganisms, an approach referred to as metabolic engineering (Nielsen, 2002).

In metabolic engineering the 'metabolic design problem' may be approached through the construction of a mathematical or in silico model of the metabolic network in question. This model can then be used to 'design' an improved metabolic network by suggesting changes in the genotype of microorganism, and techniques from modern molecular biology may subsequently be used to realize these new 'designs' and perform an experimental evaluation. Despite the promising use of mathematical models of metabolic networks in metabolic engineering, these models do, however, not only find application in industrial biotechnology as they represent a platform for in silico biology and hence play a role in medical and life science applications.

The complex nature of cellular metabolism and regulation often poses difficulties in metabolic engineering, e.g. when the flux through a specific pathway needs to be increased, and it is therefore often necessary to analyze the metabolism as a whole. The availability of complete genome sequences for several microorganisms has opened an opportunity to develop metabolic models on a genomic scale. Furthermore, the progress in experimental biology has shifted the focus of modern biology from traditional 'local' reductionist approach to 'global' holistic perspective of the cellular processes. This has resulted in the establishment of very large experimental data-

bases, or so-called omics databases, and much information on different microorganisms is therefore becoming available. A key question, however, still is to extract all the relevant information from these datasets and employ it for designing efficient industrial processes. Hence, the major challenge for metabolic engineering in this post-genomic era is to broaden its designing methodologies to incorporate the biological data on whole genome-scale. Genome-scale models represent a first step in this direction and as the reliability and accuracy of predictions/hypotheses generated by a model is directly linked to the nature and level of abstraction in the model used, it is important to constantly evaluate the value of applying genome-scale models in metabolic engineering.

In this review, we focus on the value of applying genome-scale microbial models in metabolic engineering. We also address the issue of using various genome-scale omics data sets for extracting information that can be applied in metabolic engineering.

2.3 Modeling of metabolic networks

The models that are widely used in metabolic engineering can broadly be grouped into two classes, viz., stoichiometric models and kinetic models (Nielsen, 2002;Gombert and Nielsen, 2000;Wiechert, 2002). Stoichiometric models describe the metabolic network as a set of stoichiometric equations representing the biochemical reactions in the system (Fig. 2.1). At steady state, mass balance constraints on the metabolite pools in the system can be used to determine the intracellular metabolic fluxes. The model is often represented as a stoichiometric matrix with the elements representing stoichiometric coefficient of the different metabolites in the metabolic network. The corresponding equation system is often underdetermined and analysis of the model requires imposing additional constraints (Vallino and Stephanopoulos, 2000;Stephanopoulos, 1999). In metabolic flux analysis (MFA) a number of exchange fluxes are measured to render a determined equation system. The former approach can also be effectively combined with additional information supplied by measurement of labeling pattern of certain metabolites, and is often referred to as metabolic network analysis (MNA) (Christensen and Nielsen, 2000;Klapa et al., 2003;Van Winden et al., 2003). In predictive studies, a specified objective function can be employed to determine an optimum flux distribution via linear programming, a methodology known as flux balance analysis (FBA) (varma et al., 1993;Kauffman et al., 2003). In contrast to these approaches, where a particular flux solution is sought for, one may also analyze the topology of the metabolic network through the use of convex analysis and the concepts of Elementary Flux Modes or Extreme Pathways (Schuster et al., 2000;Schilling et al., 2000;Klamt and Stelling, 2003;Palsson et al., 2003).

Kinetic models, on the other hand, define the metabolic system by combining kinetics information about specific cellular process with known stoichiometry (Gombert and Nielsen, 2000;Cronwright et al., 2002;Prathumpai et al., 2003). Thus in principle kinetic models capture the dynamic properties of the metabolic network, but a major problem associated with setting up kinetic models are the lack of kinetic data and the difference between in vivo and in vitro kinetic parameters (Gombert and Nielsen, 2000).

2.4 Current status of genome-scale metabolic models

Attempts to use genomic information in building metabolic models has resulted in genome-scale models of the metabolic networks operating in *Escherichia coli* (Edwards and Palsson, 2000;Reed et al., 2003), *Haemophilus influenzae* (Edwards and Palsson, 1999), *Helicobacter pylori* (Schilling et al., 2002), and *Saccharomyces cerevisiae* (Forster et al., 2003a). These models have been built using biological information from several sources (see Fig. 2.1) and incorporate the majority of metabolic reactions occurring in the metabolism of the studied microorganisms. These models are essentially stoichiometric models and do not explicitly contain kinetic information, regulatory mechanisms and other cellular processes. The main reason behind this is lack of information and incomplete understanding of complex cellular regulation. Thus it appears that we are still far from the goal of a complete in silico representation of the cellular metabolism. However, the stoichiometric genome-scale models themselves present a challenging task to extract, understand and use all the information contained in them. Indeed several predictions obtained from the analysis of these models have shown some promise in connecting genotype with phenotype (Price et al., 2003) and on the other end it has also been shown that the network structure (e.g. as defined by the stoichiometric model) can be used to deduce some regulatory information (Stelling et al., 2002;Beard et al., 2002;Schuster et al., 2002b). In this section we will briefly review some of the results that are of interest from metabolic engineering perspective.

2.4.1 Elucidation of design objectives of microbial metabolic networks and predicting optimal phenotypic behavior

Experimental and in silico studies with *Escherichia coli* (Edwards et al., 2001;Ibarra et al., 2002) has demonstrated that metabolic network of *Escherichia coli* has been evolved for optimization of the specific growth rate. Similar results were obtained for the growth of *Saccharomyces cerevisiae* under glucose limited conditions (Famili et al., 2003) and in both studies, a good correspondence was observed between experimental measurements and in silico predictions of substrate uptake rates, metabolite secretion rates and cell growth rates.

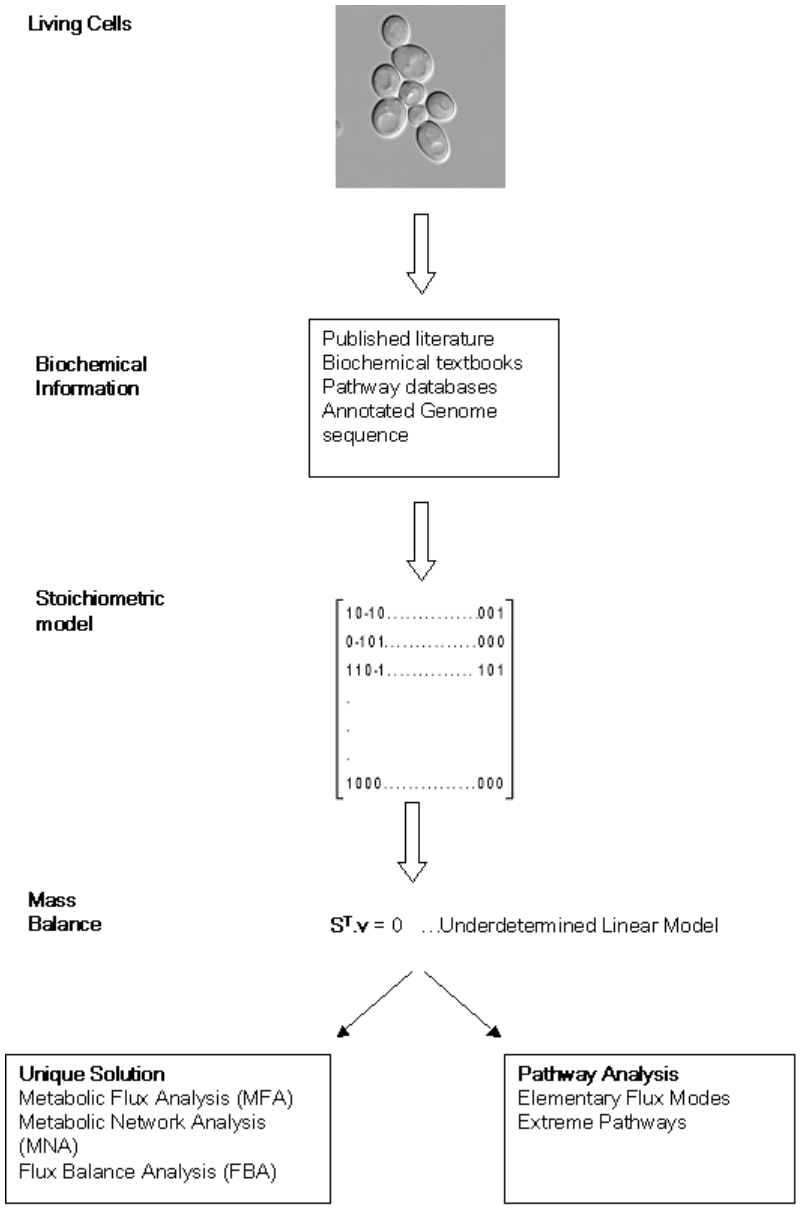


Figure 2.1. Stoichiometric modeling of metabolic networks. The knowledge about presence and stoichiometry of metabolic reactions in a particular microorganism may be extracted from various information sources (e.g. annotated genome information, biochemical text-books, published literature and pathway databases). This stoichiometric information is then summarized in a stoichiometric matrix S and a mass balance is set up under a steady-state assumption. This results in an underdetermined linear model, which can be analyzed by two different approaches. (a) The first approach, where a unique solution for the model is sought for under the given environment, the solution is obtained by constraining and/or determining some of the fluxes. In Metabolic Flux Analysis a number of exchange fluxes are measured to render a determined equation system, while in Metabolic Network Analysis in addition of constraining some fluxes, further information generated by measurement of labeling pattern of certain metabolites is used to determine the unique solution. Flux Balance Analysis uses linear optimization to determine the optimum solution with a defined objective function. (b) In the second approach, instead of looking for a single solution to the model, all possible steady-state solutions are enumerated via so called Elementary Flux Modes or Extreme Pathways using convex analysis.

In case of *Saccharomyces cerevisiae* the computed P/O ratio and energy maintenance requirements were also found to be quantitatively in agreement with experimental results. The work of Burgard and Maranas (Burgard and Maranas, 2003) provides a mathematical framework for testing whether the hypothesized metabolic objective function is consistent with experimentally investigated flux data and they show that the flux data obtained by isotopomer studies for both aerobic and anaerobic cultivations of *Escherichia coli* supports the hypothesis of optimal growth rate.

2.4.2 Predicting outcomes of genetic manipulation

Though the assumption of optimality in wild type cells may be justifiable (Edwards et al., 2001;Ibarra et al., 2002;Famili et al., 2003), Segre and coworkers (Segre et al., 2002) have demonstrated that this is not necessarily the case with genetically engineered strains which have not been exposed to long term evolutionary pressures and the flux distribution and phenotypic behavior of genetically engineered strains can be better explained through the hypothesis that such strains undergo minimal redistribution of fluxes with respect to the wild-type strain.. Nevertheless, gene deletion analysis using FBA and genome-scale models have showed good correlation with experimental data in predicting the essentiality of the genes for growth of the microorganism. The success rate of predictions were found to be 88 % for *Saccharomyces cerevisiae* (Forster et al., 2003b;Famili et al., 2003), 86 % for *Escherichia coli* (Edwards and Palsson, 2000;Edwards et al., 2001) and 60 % for *Helicobacter pylori* (Schilling et al., 2002). Though the success rate of gene deletion analysis is quite high, the failure in predicting the correct outcome can prove to be of particular importance in metabolic engineering as it indicates either incomplete or incorrect information and can lead to false predictions of metabolic designs. The microbial physiology can potentially play an interactive role in such cases, which will lead to an improved metabolic engineering performance of genome-scale models (Nielsen and Olsson, 2002).

A gene addition analysis using mixed-integer optimization along with FBA has demonstrated the in silico increase in amino acid production by *Escherichia coli* (Burgard and Maranas, 2001) using a stoichiometric model (Pramanik and Keasling, 1997). Application of genetic perturbation analysis in metabolic engineering has already been demonstrated by Nissen and coworkers (Nissen et al., 1997) by using MFA.

2.5 Improving the predictions

Beard and co-workers (Beard et al., 2002) have used Energy Balance Analysis (EBA) which eliminates the thermodynamically infeasible results from FBA. Analysis of an *Escherichia coli* genome-scale model using EBA and FBA together resulted in the same optimal specific growth rate

but different flux distributions and improved predictions over gene deletion analysis using FBA (Edwards and Palsson, 2000;Edwards et al., 2001). In an attempt to incorporate regulatory information into genome-scale stoichiometric models, Covert and coworkers (Covert et al., 2001;Covert and Palsson, 2002) introduced additional constraints through the use of Boolean operators and it has been also shown that it results in elimination of a large number of extreme pathways that are otherwise found (Covert and Palsson, 2003). In a complementary approach, a dynamic FBA problem was solved for diauxic shift in *Escherichia coli*, and qualitative agreement with experimental data was observed and the authors point out that such formulation can potentially be used to simulate a system in a dynamic environment (Mahadevan et al., 2002).

2.6 Using pathway analysis

The use of pathway analysis for successful redirection of metabolic fluxes from carbohydrate metabolism to biosynthesis of aromatic compounds in *Escherichia coli* (Liao et al., 1996) presents one of the early examples of metabolic engineering using pathway analysis. Recently Elementary Flux Mode analysis has been applied to a genetically engineered strain of *Saccharomyces cerevisiae* producing biodegradable plastic poly-β-hydroxybutyrate (Carlson et al., 2002). These examples demonstrate the use of pathway analysis in predicting the results of network/environment modification *in silico*. Pathway analysis offers possibilities for assessing structural and functional properties of metabolic networks by enumerating all possible solutions defined by the stoichiometry (Schuster et al., 2000;Schilling et al., 2000), however, use of pathway analysis approach in genome-scale models is essentially limited due to the combinatorial complexity of large metabolic networks (Klamt and Stelling, 2002).

2.7 Metabolic engineering potential of genome-scale models

The ability of genome-scale models to predict phenotypic changes resulting from genetic modifications offers a basis for rational selection of genetic targets for successful metabolic engineering and can for instance be used for hypothesis testing, that is, evaluation of proposed genetic modifications. While it seems that many "wild-type" organisms have evolved to optimize their growth yield (Edwards et al., 2001;Ibarra et al., 2002;Famili et al., 2003;Burgard and Maranas, 2003), the objective of a metabolic engineer will typically be different, for example maximum production of a desired compound. The above-mentioned tools may here be used to assess the maximum theoretical yield for the bioconversion in question or to search for reactions in a database that would increase the yield further (Burgard and Maranas, 2003). Even though some insight can be obtained from the optimal routes through the metabolic network, it should be emphasized that

these predictions does not necessarily point at the required targets for genetic modification. Evolutionary strategies have been suggested as a means to obtain the predicted properties but it should be recognized that while it is easy to select for optimal growth, it would be much harder to design experiments that will select for optimal production (see Chapter 8 for discussion on some strategies to overcome this difficulty).

For some products, it can be exploited that there is coupling between optimal growth and enhanced production, typically for by-products from the primary metabolism. One can then, while using cellular growth as the objective and observing the effect on the desired production rate, quickly screen through suitable genes to delete from the host organism or genes/reactions to add from a database. This strategy was recently applied to identify novel targets for increasing ethanol yield and xylose uptake rate in *Saccharomyces cerevisiae* (Forster, 2003). As there may be multiple optima with the same growth yield, it may be appropriate to verify both lower and upper bounds for the production rate in question, which was exemplified in a case study of succinate production in *Aspergillus niger* (David et al., 2003). It has also been shown that the nested problem of searching for gene deletions leading to optimal production while fulfilling optimal growth can be re-formulated as a single-level optimization problem (Burgard et al., 2003).

2.8 Use of genome-scale 'omics' data

The availability of genome-scale transcriptomics, proteomics and metabolomics data offer new challenges to use this data, along with physiological studies, for metabolic engineering and functional genomics of microorganisms (Nielsen and Olsson, 2002;Sanford et al., 2002;Phelps et al., 2002). To effectively use the data from different levels of cellular processes it is necessary to develop appropriate model structures and algorithms which allow examination of the global structure and properties of the metabolism while still be able to analyze at the molecular level where actual target of metabolic engineering will be identified (Ideker and Lauffenburger, 2003). In this context, genome-scale models representing the knowledge on metabolic networks can play an important role. For example in interpretation of gene expression data (Hanisch et al., 2002;Kuffner et al., 2000) and also as an complement to other data sets containing physical and functional interactions (Ideker et al., 2002).

An algorithm for combining metabolomics data with pathway analysis, which can be used to assign function to genes with unknown function, was suggested by Forster and coworkers (Forster et al., 2002). Åkesson and coworkers (Akesson et al., 2004) have demonstrated that incorporating even the simplest form of regulatory information, namely the presence or absence of a particular

gene, from gene expression data improves the predictions of flux distributions obtained from genome-scale model of *Saccharomyces cerevisiae*. Bro et al. (Bro, 2003) demonstrated a successful application of genome wide expression data where the galactose uptake rate in *Saccharomyces cerevisiae* was increased by 70 %. This work also demonstrates and highlights an important issue of using known pathway information while analyzing the genome-wide data, which when used alone, can pose difficulties for using statistical analysis in identifying metabolic engineering targets.

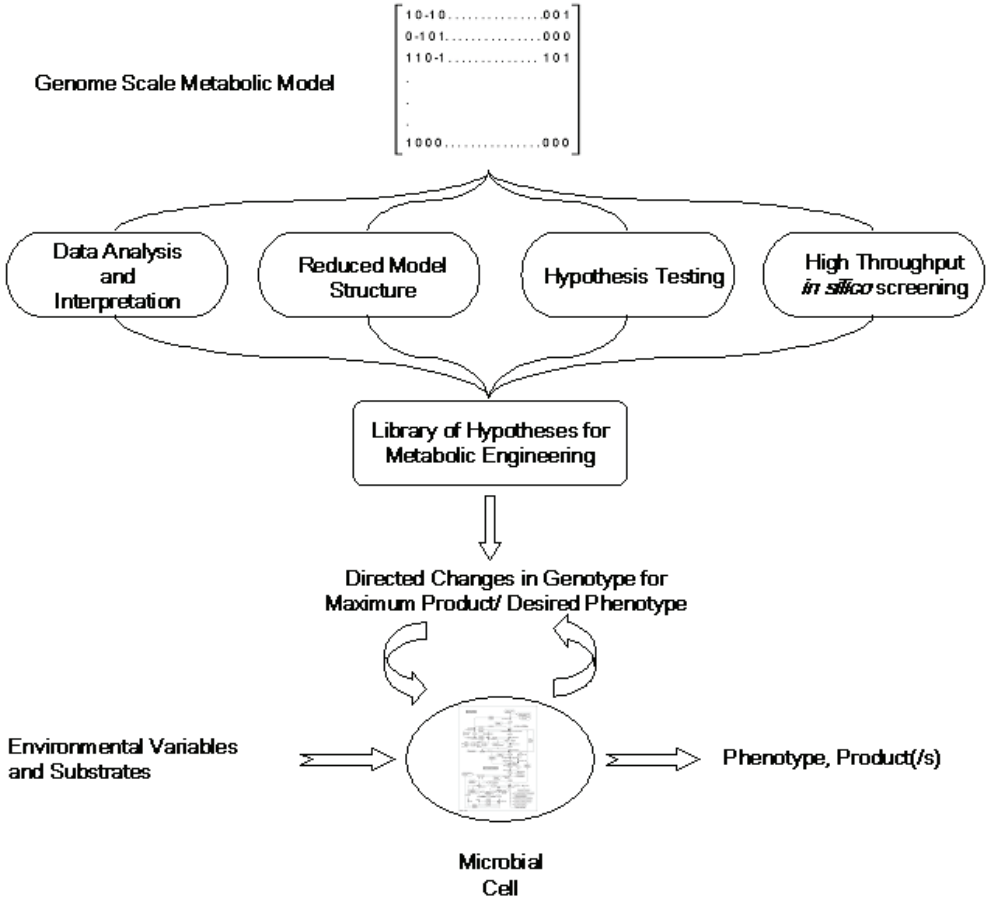


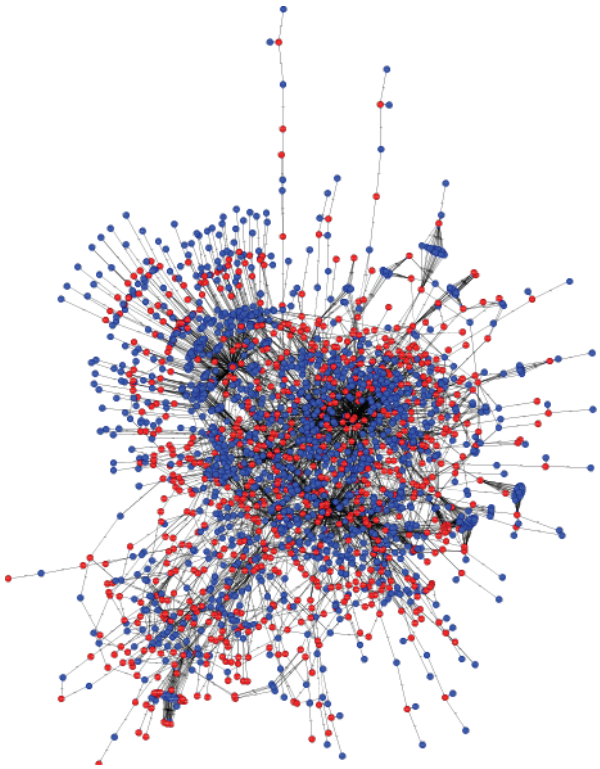
Figure 2.2. Applications of genome-scale models in metabolic engineering. Genome-scale models can serve to test and score the biochemical hypotheses about genetic manipulation in the metabolism. In combination with pathway databases, they may also be used for high throughput in silico screening of a large number of gene insertion and/or deletion mutants. The metabolic networks incorporated in the genome-scale models can also be useful for interpretation of experimental results, for instance in integrative analysis of omics data or serving as templates for simplified models to be used in e.g. metabolic network analysis.

2.9 Conclusions

In spite of the limitations on kinetics information of many cellular processes, presently available genome-scale models can play an important role in metabolic engineering of microbial cells. Figure 2.2 illustrates some key potential uses of current genome-scale models in metabolic engineering. The power of genome-scale models to predict outcomes of genetic manipulation can be used to score the biochemical hypotheses for manipulation of genotype. Alternatively, a large number of gene addition and/or deletion mutants can be screened in silico to generate and score a library of hypotheses. The significance of such scores/hypotheses will be highly dependent on the reliability and gravity of the model but nevertheless this can serve as an important guideline for prioritization of hypotheses to be tested in vivo. In addition to the design problem, the models may also be important for analysis and interpretation of omics data as discussed above. The metabolic networks embedded in the models can also be used as a reference/parent model from which a reduced model is generated depending on the context of problem, and this model can then be used, e.g. in the analysis of 13C labeling experiments or metabolic design. This could be of interest since the complexity of genome-scale models can sometimes pose difficulties in analysis of a specific problem.

Chapter 3: Uncovering transcriptional regulation of metabolism by using metabolic network topology

This chapter is based on the publication:
Patil, K. R. & Nielsen, J. Uncovering transcriptional regulation of metabolism by using metabolic network topology. PNAS 102, 2685-2689 (2005).



"It seems very pretty,' she said when she had finished it, `but it's RATHER hard to understand!" (You see she didn't like to confess, ever to herself, that she couldn't make it out at all.) `Somehow it seems to fill my head with ideas -- only I don't exactly know what they are!"

3.1 Abstract

Cellular response to genetic and environmental perturbations is often reflected and/or mediated through changes in the metabolism, since the later plays a key role in providing Gibbs free energy and precursors for biosynthesis. Such metabolic changes are often exerted through transcriptional changes induced by complex regulatory mechanisms coordinating the activity of different metabolic pathways. It is difficult to map such global transcriptional responses by using traditional methods, as many genes in the metabolic network have relatively small changes at their transcription level. We therefore developed a novel algorithm that is based on hypothesis driven data analysis to uncover the transcriptional regulatory architecture of metabolic networks. By using information on the metabolic network topology from genome-scale metabolic reconstruction, we show that it is possible to reveal patterns in the metabolic network that follow a common transcriptional response. Thus, the algorithm enables identification of so-called *reporter metabolites* (metabolites around which the most significant transcriptional changes occur) and a set of connected genes with significant and coordinated response to genetic or environmental perturbations. We find that cells respond to perturbations by changing the expression pattern of several genes involved in the specific part(s) of the metabolism where a perturbation is introduced. These changes are then propagated through the metabolic network, due to the highly connected nature of those.

3.2 Introduction

Linking the genome to its functioning metabolism is of substantial interest not only in studying human diseases (Peltonen and McKusick, 2001) but also for identifying metabolic engineering targets in biotechnological applications (Patil et al., 2004) (Chapter 2)(Nielsen and Olsson, 2002). Transcriptional analysis represents a high-throughput and genome-wide approach for linking the set of expressed genes to functional metabolism of the cell. Indeed, several studies using genome-wide gene expression analysis have shown that the transcriptional regulation plays an important role in regulating metabolism in response to perturbations (Ihmels et al., 2004;DeRisi et al., 1997;Miki et al., 2001). Although many statistical methods and clustering algorithms provide tools to analyse such transcriptomics data (Eisen et al., 1998;Herrero et al., 2004;Sherlock, 2000), these methods seldom provide insight into the regulatory architecture of the metabolic networks without intelligent analysis of the results (up-down regulation of genes of interest or correlation between genes of interest). This is primarily due to the hypothesis that there may be *all to all* interactions amongst the genes being analysed, resulting into many biologically nonsignificant results. One of the ways to address this problem is to integrate known biological interactions, e.g. pro-

tein-protein interactions, in the analysis of transcription data (Ideker et al., 2001). Such an approach essentially reduces the degrees of freedom in data analysis using knowledge of molecular interactions occurring in the cell. The organization and functioning of the cell can be viewed as a complex network of molecular interactions. These interactions are mediated not only by physical contacts between individual molecules (e.g. protein-protein and protein-DNA interactions), but also result from the functional coupling of certain molecules or groups of molecules (Lee et al., 2004). Cellular metabolism can thus also be viewed as a network of functional interactions between enzymes and metabolites. This metabolic network represents the channels for the flow of material and generation of Gibbs free energy, which are constrained by the conservation laws of mass and energy. Consequently, we hypothesized that the topology of the interactions involved in metabolism can be used to understand the underlying regulatory mechanisms (e.g. at transcriptional level) controlling this flow of mass and energy. To test this hypothesis, we developed a novel algorithm that integrates gene expression data with topological information from genome-scale metabolic models. This enabled systematic identification of so-called *reporter metabolites* that represent hot spots in terms of metabolic regulation. This is one of the first attempts to infer the global role of a metabolite based on mRNA expression patterns and metabolic stoichiometry without direct measurement of metabolite concentration. The algorithm also identifies the significantly correlated metabolic subnetworks following direct or indirect perturbations of the metabolism.

3.3 Algorithm

Figure 3.1 schematically illustrates the proposed algorithm, which is described step by step in the following.

3.3.1 Graph-theoretical representation of the metabolic network

The complete metabolic network in the cell can be represented as a bipartite undirected graph, here referred to as a metabolic graph (Figure 3.1) (Supplementary material). In this metabolic graph, metabolites as well as enzymes are represented as nodes and interactions between them are represented as edges. Thus a metabolite node is connected to all the enzyme nodes that catalyse a reaction involving that particular metabolite, and an enzyme node is connected to all the metabolites that take part in the corresponding reaction. This graph is bi-partite since neither metabolite nor enzyme nodes are directly connected amongst them.

We also define a unipartite undirected graph, here referred to as enzyme (or reaction) interaction graph (Figure 3.1) (Supplementary material). In this graph only enzymes are represented as the

nodes, and the two enzymes sharing a common substrate in the corresponding reactions are connected to each other. Thus edges in this graph represent the metabolites shared by two enzymes. Some enzymes catalyse several different reactions, and these enzymes are represented by a single node. This node is linked to all enzyme nodes that are connected to the different reactions carried out by this enzyme.

3.3.2 Mapping and scoring of transcription data

The transcriptional data used in this study can be classified into two categories. The first category includes data where two different strains (or conditions) are compared and with multiple measurements for each strain (or condition). We refer to this data type as the differential data. The second category of data is multidimensional data, e.g. gene expression measured over a time course or with analysis of multiple strains, with or without multiple measurements at the same time point or strain.

Differential data can be mapped on the enzyme nodes of the metabolic or enzyme-interaction graph with a specification of the significance of differential gene expression. Here we used the student's t-test to obtain p-values, with p_i representing the significance of the change for each enzyme. Each p_i can subsequently converted to a Z-score of the enzyme node (Z_{ni}) using the inverse normal cumulative distribution (CDF, θ^{-1}).

$$Z_{ni} = \theta^{-1}(1 - p_i)$$

In the case of multidimensional data, the absolute Pearson correlation coefficient, P_j is calculated between all pairs of nodes (enzymes) connected by an edge in the enzyme-interaction graph. The P_j of an edge can be converted to a Z-score for that edge (Z_{ej}) using the normal CDF.

$$Z_{ej} = \theta^{-1}(P_j)$$

The Z-score follows a standard normal distribution for a random data, where p-values or Pearson coefficients follow a uniform distribution.

3.3.3 Method for identification of *Reporter Metabolites*

In order to identify the reporter metabolites each metabolite node in the metabolic graph is scored based on the normalized transcriptional response of its neighbouring enzymes. In case of differential data, the normalized transcriptional response was calculated as size-independent aggregated Z-scores of the k neighbouring enzymes.

$$Z_{metabolite} = \frac{1}{\sqrt{k}} \sum Z_{ni}$$

$Z_{metabolite}$ scores can be corrected for the background distribution by subtracting the mean (μ_k) and dividing by the standard deviation (σ_k), of the aggregated Z-scores of several sets of k enzymes chosen randomly from the metabolic graph.

$$Z_{metabolite}^{corrected} = \frac{(Z_{metabolite} - \mu_k)}{\sigma_k}$$

For multidimensional data, the neighbouring enzymes of a metabolite in the metabolic graph are represented as enzyme interaction graph with all enzymes connected to each other and hereby Z-scores for each edge (Z_{ej}) can be calculated as described before. Subsequently $Z_{metabolite}$ score can be calculated and corrected for the background distribution in the same way as for differential data.

The scoring used for identifying *reporter metabolites* is basically a test for the null hypothesis, "neighbour enzymes of a metabolite in the metabolic graph show the observed normalized transcriptional response by chance". The metabolites with the highest score (typically up to 10) are defined as *reporter metabolites*, and these mark spots in the metabolism where there is substantial regulation either to maintain homeostasis, i.e. a constant level of the metabolite, or to adjust the concentration of the metabolite to another level required for proper functioning of the metabolic network.

3.3.4 Method for identification of highly correlated subnetworks

As the next step in uncovering the transcriptionaly correlated parts of the metabolism following a perturbation, we addressed the problem of identifying highly correlated connected sub graphs (subnetworks) within the enzyme interaction graph. Firstly we define the score Z_s of a connected subnetwork, which characterizes the biological activity, or the aggregate transcriptional response of subnetwork as:

$$Z_s = \frac{1}{\sqrt{k}} \sum_{n/e \in s} Z_{ni/ej}$$

We used the Z-score of the node, Z_{ni}, in case of differential data and Z-score of the edge, Z_{ej}, in case of multi-dimensional data. As in case of the *reporter metabolites*, we corrected the Z_s score for the background distribution of the subnetworks of the same size, randomly sampled from the same network.

Finding the subnetwork with the highest score is a NP-hard problem and was approached by using a simulated annealing algorithm (Ideker et al., 2002) (see Supplementary material for the details of the implemented algorithm). Within the identified subnetwork further subnetworks may be searched by repetition of the algorithm over the subnetwork previously obtained (sub-networks reported in Table 3.2 and Supplementary Table S-3.3 were obtained after applying the simulated annealing to larger subnetworks resulting from analysis of the whole metabolic network). We also note that simulated annealing is a stochastic method and does not guarantee that the global optimal solution is found. Moreover the resulting subnetwork solution might differ depending on the initial conditions and the parameters. We addressed these problems by repeating the simulated annealing search several times (about 10) and selecting the subnetwork with the highest score. We observed that it was possible to obtain robust solutions with high scores and biological significance by optimizing the parameters of simulated annealing.

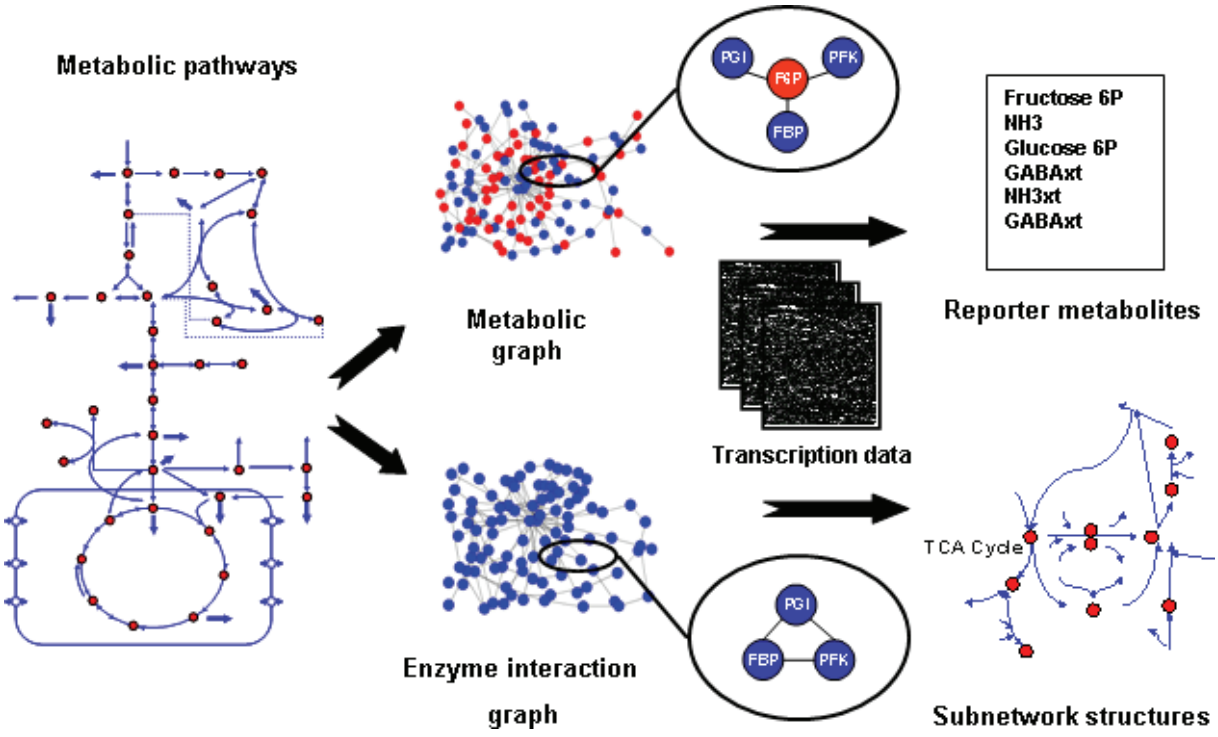


Figure 3.1. Illustration of the proposed algorithm for identifying *reporter metabolites* and subnetwork structures signifying transcriptional regulated modules. A metabolic network (set of reactions) is converted to a bipartite graph (metabolic graph) and a unipartite graph (enzyme interaction graph) representation. Gene expression data from a particular experiment is then used to identify highly regulated metabolites (*reporter metabolites*) and significantly correlated subnetworks in the enzyme interaction graph.

3.4 Results

We implemented the algorithm for analysis of transcription data from the yeast *Saccharomyces cerevisiae*. Besides its use as a cell factory, this yeast is extensively used as a model system for studying human diseases (Botstein et al., 1997). We used the recently reconstructed genome-scale metabolic network of *S. cerevisiae* (Forster et al., 2003a) to generate the metabolic and the reaction interaction graphs, and subsequently applied the algorithm to many yeast gene expression datasets to illustrate the algorithm.

3.4.1 Deletion of a gene encoding an enzyme

We first analysed transcription data from a wild-type strain of *S. cerevisiae* and a mutant with deletion of the gene *GDH1*, which encodes for NADPH dependent glutamate dehydrogenase – an enzyme that plays an important role in ammonia assimilation. Physiological analysis of this strain demonstrated an effect on redox metabolism, as observed through increased ethanol yield and decreased glycerol yield (Bro et al., 2004). However, conventional transcriptome analysis of this mutant, where differentially expressed genes are identified using a statistical test (e.g. t-test analysis with Bonferroni correction) did not enable identification of the overall effect of this genetic perturbation on the metabolism. Despite these results, using our novel algorithm we identified several key *reporter metabolites*, which include: ammonia, glucose 6-phosphate, fructose 6-phosphate and sedoheptulose 7-phosphate (Table 3.1). The fact that ammonia (both intracellular and extra-cellular ammonia) is identified as a *reporter metabolite* is biologically reasonable, as ammonia assimilation has been altered. It may intuitively be more difficult to understand why the three sugar phosphates appear as *reporter metabolites*. However, these three metabolites represent branch points between the Embden Meyerhof Parnas (EMP) pathway and the pentose phosphate (PP) pathway. Upon deletion of *GDH1* the requirement for NADPH in connection with cellular growth is reduced by more than 40% (Nissen et al., 2000), and this reduces the requirement for shunting glucose through the pentose phosphate pathway, which acts as the primary source for NADPH in *S. cerevisiae*.

Looking at the highly correlated metabolic subnetwork we found the high scoring subnetwork to consist of 181 genes distributed in 68 MIPS functional categories (Mewes et al., 2004) (Supplementary Figure S-3.2), of which 31% belong to MIPS functional categories amino acid metabolism and transport, carbohydrate utilization and nucleotide metabolism. Further analysis of the 181 genes subnetwork resulted in identification of a 34 genes subnetwork (Table 3.2). This subnetwork consists of 10 genes (apart from *GDH1*) encoding enzymes catalysing oxido-reductive reactions involving the co-factors NADPH/NADH, clearly demonstrating the effect of *GDH1*

deletion on redox metabolism. In fact, these co-factors represent the main links in this subnetwork, which involves two key nodes in the cellular metabolism (Fig. 3.2): 1) the node between the EMP pathway and the PP pathway and 2) the node around α-ketoglutarate. The first node is known to be controlled by the requirement for NADPH. The decrease in expression of genes of the PP pathway is consistent with a decreased flux through this pathway in a similar mutant (Moreira dos et al., 2003). The second node is directly perturbed, and it makes sense that this results in a transcriptional response of enzymes around this node. It has indeed been shown that in a *gdh1Δ* mutant the level of α-ketoglutarate is increased (DeLuna et al., 2001), and this is consistent with a decreased expression of the genes *KGD* and *LSC*, both encoding enzymes downstream of α-ketoglutarate.

3.4.2 Deletion of a gene encoding regulatory protein

In order to further evaluate the method, we also analysed transcription data from a *grr1Δ* mutant of *S. cerevisiae* compared with a wild-type strain, both grown at high glucose concentrations (Westergaard et al., 2004). Grr1p is a ubiquitin-protein ligase that plays a role in glucose repression (Flick et al., 2003). Overall it is known that Grr1p deactivates the Rgt1p transcriptional repression of several hexose transporters, and the important role of Grr1p in regulating sugar transporters is clearly seen from the list of *reporter metabolites* identified in this case (Table 3.1). Among the 10 most important *reporter metabolites* 6 are hexoses, all transported by the group of *HXT*-genes in *S. cerevisiae*. The other *reporter metabolites* include glutamine, orthophosphate and glycogen. Glutamine is playing a key role in the nitrogen metabolism, which is normally considered also to be regulated by Grr1p, even though a direct link has not been established (Bernard and Andre, 2001). Orthophosphate is involved in a large number of reactions in the central carbon metabolism, and the identification of this *reporter metabolite* is a clear indication of the multitude of effects caused by deletion of *GRR1*. In the *grr1Δ* mutant, a high scoring metabolic subnetwork of 204 genes was identified, and further analysis of this network resulted in identification of a 52 genes subnetwork (Table 3.2). Besides several genes encoding sugar and amino acid transporters that are known to be regulated by Grr1p, this subnetwork also contains many other genes involved in amino acid metabolism.

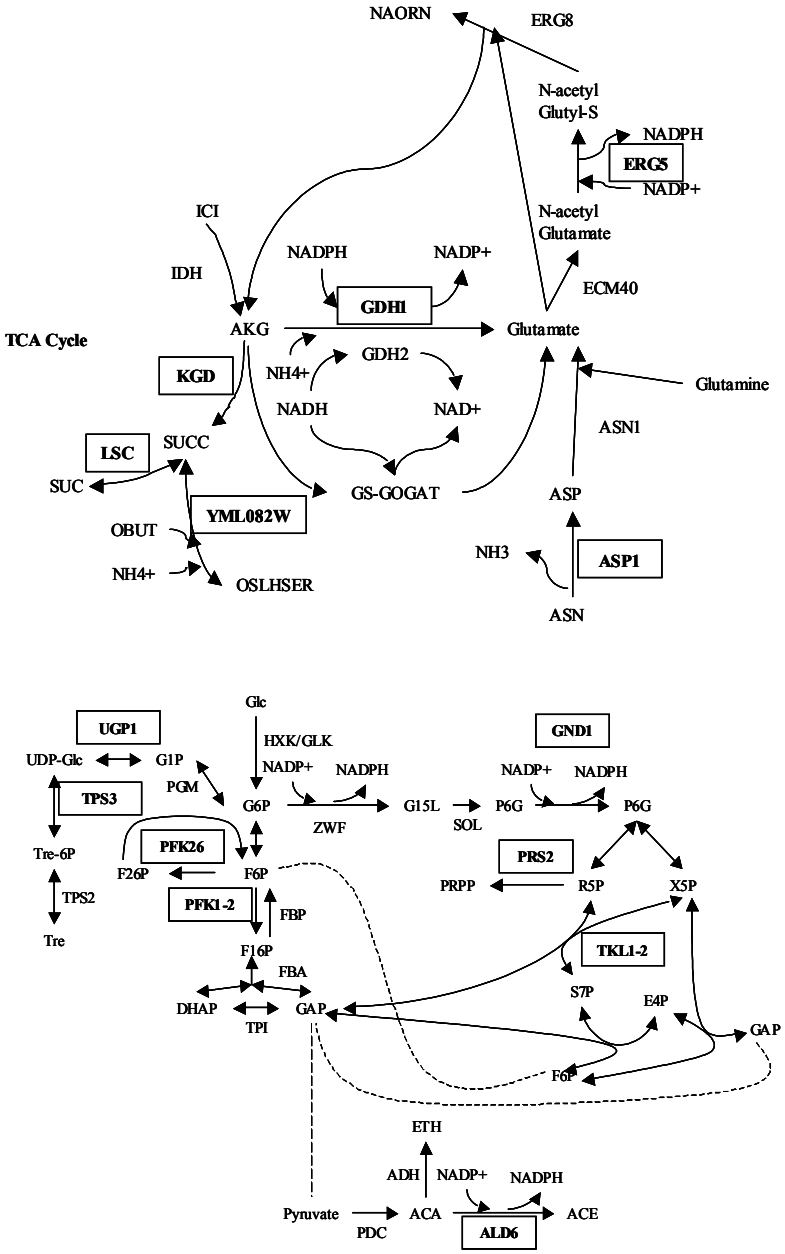


Fig. 3.2. Parts of *Saccharomyces cerevisiae* metabolism that are represented in the subnetwork identified for the *gdh1Δ* dataset. Genes present in the subnetwork are given in boxes.

3.4.3 Multi-dimensional data

To illustrate the application of the method for analysis of transcription data measured over several different environmental conditions, we analyzed transcription data for *S. cerevisiae* grown on four different carbon sources, glucose (a hexose), maltose (a disaccharide), and two C-2 compounds, ethanol and acetate(Daran-Lapujade et al., 2004). For analysis of this type of dataset it is intuitively more difficult to interpret the results in terms of the changes in physiology as the data spans a multi-dimensional space. However, the *reporter metabolites* (Table 3.1) still reflect the meta-

bolic reprogramming in response to the changes in carbon source. Maltose is an obvious *reporter metabolite* as enzymes involved in uptake and metabolism of this sugar are induced only in the presence of maltose. The presence of glyoxylate and carnitine as *reporter metabolites* is due to the key roles of these metabolites during growth on C-2 compounds (Gancedo and Gancedo, 1997). Appearance of H^+ as a *reporter metabolite* illustrates the ability of the algorithm to identify metabolites indirectly involved in metabolism, as transport of maltose and acetate is coupled with proton transport across the cell membrane. We also performed a pair wise comparison of the four carbon sources and the results are provided in the Supplementary material.

3.5 Large-scale *reporter metabolite* analysis

To further evaluate the algorithm, we performed *reporter metabolite* analysis of about 47 transcriptional datasets (Supplementary material). In all these cases, *reporter metabolites* provided useful information about the metabolic changes underlying the particular experiment, e.g., the *reporter metabolites* identified for the comparison of carbon and nitrogen limited conditions clearly show the underlying metabolic changes in major pathways for utilization of these substrates. We also found that relatively few metabolites were identified as *reporter metabolites* for many of the conditions analysed. This is due to the fact that similar types of perturbations are introduced in many of these studies (e.g. change in substrate, comparison between aerobic and anaerobic conditions). Nevertheless it is interesting to note that one-third of the all the metabolites in the metabolic graph were identified as a *reporter metabolite* in at least one of the studies (see Supplementary material for the complete distribution). Moreover, the average rank of any metabolite (defined as arithmetic average of ranks of a metabolite, based on $Z_{metabolite}$ score, from all conditions analysed) was found to be more than 150, further illustrating the uniqueness of *reporter metabolites* for the particular experiment.

Table 3.1. Reporter metabolites (metabolites with highest neighbor subnetwork scores) for *gdh1Δ*, *grr1Δ* and carbon-sources datasets.

gdh1Δ			**grr1Δ**			**Carbon source**		
Fructose 6-phosphate	15	5	L-Glutamine	20	5	Maltose	4	1
Glucose 6-phosphate	11	6	GLCxt	14	-	Carnitine	3	1
NH3xt	3	-	Mannose	15	2	(R)-Pantoate	2	1
NH3	32	7	Fructose	14	3	Glyoxylate	6	5
GABAxt	2	-	FRUxt	12	-	6P-gluconate	5	1
CTP	8	1	Glycogen	4	-	Episterol	2	-
Fructose 1,6-bisphosphate	4	4	Orthophosphate	65	3	3-Demethylubiquinone-9M	2	-
Sedoheptulose 7-phosphate			Glucose	28	6	H^+EXT	42	-
CO2M	5	2	MANxt	11	-	3-Phosphonooxypyruvate	3	1
N-Acetyl-L-glutamate 5-semialdehydeM	12	-	Homocitric acid	2	2	1-Phosphatidyl-1D-myo-inositol 4P	4	2
	2	-						

Only the top ten scoring metabolites are shown. The first number behind the reporter metabolite is the number of neighbors to the reporter metabolite (or the number of reactions in which the reporter metabolite participates in) and the second number is the number of KEGG pathways in which the reporter metabolite appears. The metabolite names ending with 'M' and 'xt' indicate that the metabolite is present in mitochondrial compartment and extra-cellular medium respectively. Since KEGG pathways do not classify metabolites in this fashion, the corresponding fields in the table are empty.

Table 3.2. Genes included in the subnetworks obtained by analysis of gene expression datasets for *gdh1Δ*, *grr1Δ* and carbon-sources.

gdh1Δ	**grr1Δ**	**Carbon source**
PFK2 PMP2 PFK1 QNS1 MEP2 GDH1 ADE3 PFK26 HTS1 UGP1 SAM1 BIO3 ERG6 SAH1 PCT1 PRS2 TKL1 TRP5 TPS3 GND1 ALD6 SCS7 BNA1 HOM6 PUR5 YML082W ASP1 KGD1 LEU4 LSC1 ARG5 MET13 PUT4 UGA4	HXK1 HXT3 MAL32 STL1 DIP5 YGL186C TAT2 MUP1 SHM2 ADE3 YER053C FBP1 ARO2 GLC3 ARO3 ADE6 HIS7 GUA1 RIB1 ACS2 HIS1 PFK2 YDR341C URA6 ARG1 ADE12 CPA2 PDC5 LEU2 LEU1 MDH3 YAR075W ADH5 GAD1 ASN2 MET22 SER2 GDH3 PNC1 ILV1 YMR293C LYS21 LYS20 KGD1 NDI1 RIP1 CYB2 ACH1 XKS1 PGI1 INO1 PGM2	HXK1 HXT2 ACS1 MET22 ARO2 THR4 SER1 GSH1 INM1 TOR1 PRO1 PIK1 PRS2 FUR1 QRI1 LYS20 NAT2 HMGS HMG1 PAN5 ERG3 YJR078W ERG11 ERG25 YBR006W ERG2 CAT2 CIT2 AAT2 BAT2 BAP2 SAM3 BIO2 MDH2 FDH1 GCV1 DFR1 GND2 GND1 PCK1 SOL3 NDH2 YFL030W ICL1 SFC1 MLS1 ACH1 PGM1

The subnetworks listed were obtained through simulated annealing search in a larger subnetwork (see Algorithm description and Supplementary Table S-3.2). For the subnetwork from the *gdh1Δ* dataset, bold names represent enzymes directly involved in redox metabolism.

3.6 Discussion

Reporter metabolites and corresponding subnetworks from all three cases, representing three different types of perturbations (namely deletion of a gene encoding enzyme, deletion of a gene encoding regulatory protein and change in environment of cell) clearly project the metabolic changes following these perturbations. As the transcriptional changes at individual gene levels are small, these are not identified using conventional statistical significance tests or clustering methods (Supplementary material), whereas our hypothesis driven analysis of transcription data enables identification of small and co-ordinated changes in expression levels. We also note that several of the identified *reporter metabolites* are involved in a relatively large number of reactions (Table 3.1) that are distributed in several different KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa et al., 2004) pathways. Thus, mapping of transcriptional changes onto KEGG pathways, as often done for visual representation of the transcriptional changes may be misleading.

The metabolic graph of *S. cerevisiae* consists of 2000 nodes (825 metabolites and 1175 reactions) and 4196 edges, while the reaction interaction graph has 1175 nodes and 57217 edges. Notably, a large fraction of these edges represent interactions due to energy and redox cofactors giving highly connected graphs (the average path length between any two nodes is 5.17 and 2.49 for the metabolic graph and the reaction interaction graph, respectively) with "small world properties"(Fell and Wagner, 2000;Jeong et al., 2000). The high degree of connectivity of the metabolic network implies that the disturbance at any node in the network can affect all branches of the metabolism and hence demands a global control. This can be seen from the subnetwork analysis where we found large significant subnetworks spanning all branches of the metabolism (Supplementary Table S-3.2). Such changes are, however, centred on the perturbed node (/s), as can be seen from the *reporter metabolite* analysis that identifies such nodes in the metabolism.

Owing to the high connectivity of the metabolic network, the here-reported algorithm is found to be quite robust to alterations in the metabolic graph (e.g. removal of certain metabolites). To evaluate this, we removed some of the highly connected co-factors from the graph and studied the effect on the network connectivity and subnetworks obtained for the *GDH1* dataset (Supplementary material). It was possible to obtain about 75 % overlap with the original subnetwork even after the removal of both redox co-factors (NAD+/NADH and NADP+/NADPH) and ATP/ADP pair, which resulted in 27 % reduction in the number of edges. The result was most sensitive to the removal of NADP+/NADPH, which is consistent with the fact that GDH1 encodes for a NADPH dependent enzyme. Notably, the removal of NAD+/NADH did not influ-

ence the results significantly even though it resulted in a substantial decrease in the number of edges in the network

Although the regulatory network structure defines the details of how the transcriptional regulatory program is executed, the metabolic network itself seems to guide this machinery, which we see as the consequence of the fact that metabolic regulation has been designed and evolved *for and around* the metabolites. We exploited this hypothesis by developing an effective algorithm that enables understanding the transcriptional changes of the metabolism following genetic and environmental perturbations. Apart from uncovering the architecture of the transcriptional changes following known perturbations, our approach will also be useful in identifying the effects of unknown or poorly characterized disturbances, e.g. deletion of ORF with unknown function or exposure to a drug, and hereby provide clues to the role of the ORF or the drug on the cellular metabolism.

Acknowledgements: We are grateful to I. Rocha and M. Åkesson for fruitful discussions. We also thank J. Villadsen, M. Hjortsø, G. Lettier, A. P. Oliveira, T. Grotkjær and M. Jewett for helpful suggestions.

3.7 Supplementary material

3.7.1 Supplementary methods

Graph-theoretical representation of the metabolic network

A set of metabolic reactions can be represented as a graph with metabolites and/or reactions (enzymes) as nodes and interaction between them as edges. For example:

HXK : GLC + ATP = G6P + ADP .

PGI : G6P = F6P .

PFK : F6P + ATP = F16P + ADP .

FBP : F16P = F6P .

The above set of four reactions can be visualized as a bi-partite metabolic graph and an enzyme interaction graph as shown in Supplementary figure S-3.1.

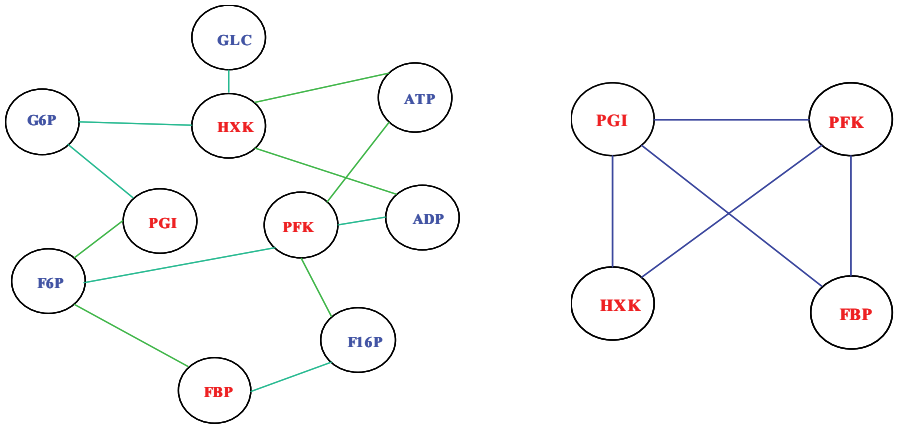


Figure S-3.1. Bi-partite and enzyme interaction representation of the metabolic network.

Simulated annealing algorithm

The algorithm used to find highly correlated subnetworks in an enzyme interaction graph is a slightly varied version of the algorithm proposed by Ideker *et al.* (Ideker et al., 2002). An enzyme interaction graph G consists of a set of nodes (enzymes) and edges (interactions due to shared metabolites). Each node in the graph G is associated with a binary variable marking it visible or invisible. G_{sub} denotes the sub graph of G induced by the visible nodes. The score of the graph G_{sub} at any iteration i, Z_s^i, is defined as the score of the highest-scoring connected component of G_{sub}.

Simulated annealing is performed from the starting temperature T_{start}, which geometrically decreases to T_{end} in N number of iterations. T_i denotes the temperature at iteration i. With these parameters defined, simulated annealing proceeds in the following steps.

Initialize G_{sub} by setting each node to visible/invisible with equal probability. Different initialisation scheme, e.g. all visible nodes, can be used depending on the nature of the problem.

For i = 1 to N DO

Randomly select a node of G and toggle its state (visible/invisible);

Compute the score Z_s^i for sub graph G_{sub};

IF ($Z_s^i > Z_s^{i-1}$) THEN keep the change, i.e. keep the selected node toggled;

ELSE keep the change with probability $p = e^{(Z_s^i - Z_s^{i-1})/T_i}$;

Quench the resulting G_{sub} at $T = 0$, to explore all adjoining possibilities so as to ensure the local maximum.

Output the high-scoring connected components of the resulting sub graph G_{sub}.

In order to improve the efficiency of the algorithm we used following set of heuristics.

Hubs (highly connected nodes) in the network tend to decrease the performance of the algorithm, as addition of such nodes changes the size of the connected components drastically. To avoid this we use the same heuristic as used by Ideker *et al.* (Ideker et al., 2002). When adding a hub to the network, all neighbours that are not part of the high-scoring component, are removed from the subnetwork (i.e. made invisible) simultaneously.

In case of multi-dimensional data, subnetworks are scored using the edge scores. Moreover, to improve the efficiency of the algorithm, we employed a two-stage simulated annealing algorithm. In the first stage, edges are toggled instead of the nodes. The resulting sub graph from the first stage is then used for the next simulated annealing stage where nodes are toggled as described before. This heuristic increased the speed and efficiency of the algorithm by an order of magnitude.

We also extended the algorithm to search for more than 1 subnetworks simultaneously (Ideker et al., 2002).

Although we searched for more than one subnetworks simultaneously, we found a single high scoring subnetwork and several small, low scoring subnetworks in all cases studied.

3.8 Supplementary Discussion

3.8.1 Distribution of subnetwork genes into different functional categories

The subnetworks obtained from the analysis of *gdh1Δ*, *grr1Δ* and carbon-sources datasets consisted of 181, 204 and 179 genes respectively (Supplementary Table S-3.2). The distribution of the genes from *gdh1Δ* subnetwork into different MIPS functional categories (Mewes et al., 2004) is shown in Supplementary Figure S-3.2. Amongst these genes, 25 % belong to the MIPS functional categories amino acid metabolism and transport, carbohydrate utilization and nucleotide metabolism. It seems natural that the amino acid metabolism is likely to be effected upon deleting the main route for ammonia assimilation. The effect on the carbon metabolism is likely to be a consequence of the interactions with the amino acid metabolism through the supply of precursor metabolites, particularly through many transamination reactions. Finally, deletion of *GDH1* will

affect the glutamate and glutamine levels in the cells, and as glutamine serves as amino-donor in the nucleotide metabolism, deletion of *GDH1* may have an indirect influence on the expression of genes involved in nucleotide metabolism. This shows that the transcriptional regulatory program of cell, following a single genetic perturbation, is global and spread across many branches of the metabolic network. Notably, we also obtained such a widespread response for the subnetworks obtained from the analysis of other two datasets.

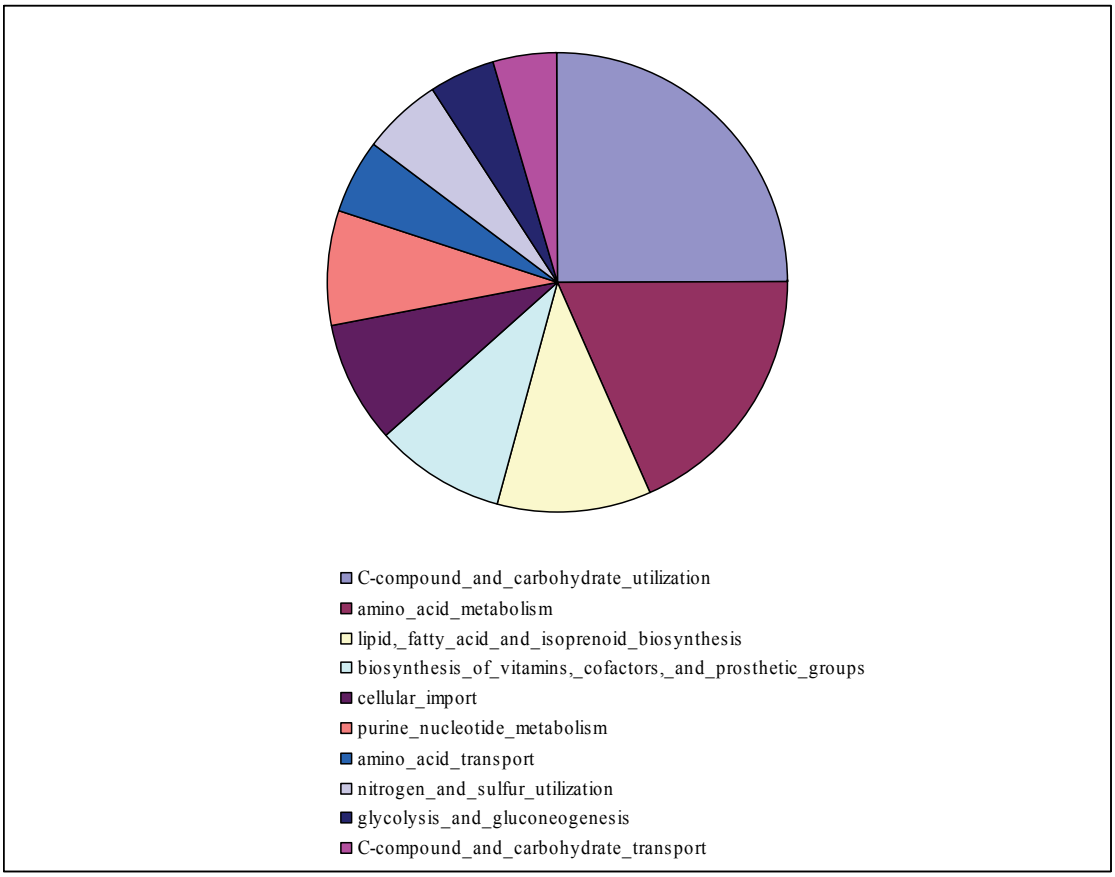


Figure S-3.2. Distribution of the genes from high scoring subnetwork for *gdh1Δ* dataset into different MIPS functional categories. For simplicity, only first ten categories with highest numbers of entries are shown.

3.8.2 Pair wise comparison of different carbon sources

In addition to the collective analysis of the carbon-sources dataset, we also performed pair wise analysis of the gene expression data for different carbon sources, with p-values calculated using t-test. The results of this analysis are summarized in Supplementary Table S-3.1 and Supplementary Table S-3.3. Reporter metabolites, together with the subnetworks, clearly identify the parts of the metabolism most affected due to a change in the carbon source.

3.8.3 Clustering analysis of the carbon sources dataset

Results of cluster analysis of the carbon-sources dataset were compared with the results obtained using here reported algorithm. Carbon sources dataset was subjected to k-means and self organizing map (SOM) clustering(Sturn et al., 2002). In order to enable comparison between clustering and subnetworks, clusters were ranked using normalized scores similar to that used for scoring subnetworks. Cluster-scores were calculated by considering a cluster as subnetwork, with all member genes connected to each other. Supplementary Table S-3.4 shows overlap between the genes present in the subnetwork, identified using our method, and high scoring clusters. We note that only about 10% of the genes within the subnetwork identified with our method are represented in a given cluster.

It is difficult to see the underlying metabolic changes, following the changes in carbon source, from the results of cluster analysis. Firstly, all the genes appear in at least one of the clusters, as opposed to the reduced set of metabolites (reporter metabolites) and genes (subnetwork) obtained with here reported algorithm. Secondly, our algorithm allows relatively weakly correlated genes to be grouped together as a subnetwork. Subnetwork may contain genes that are correlated to a lesser degree within subnetwork as compared with some of the genes outside this subnetwork, a consequence of imposing structural constraints on possible interactions. This can be seen from Supplementary Figure S-3.3. Also, it should be noted that the usual clustering methods account for direction of the changes, which could result in loss of information since metabolic genes are often negatively correlated, e.g. up-regulation of glucose transporters and down regulation of maltose transporters.

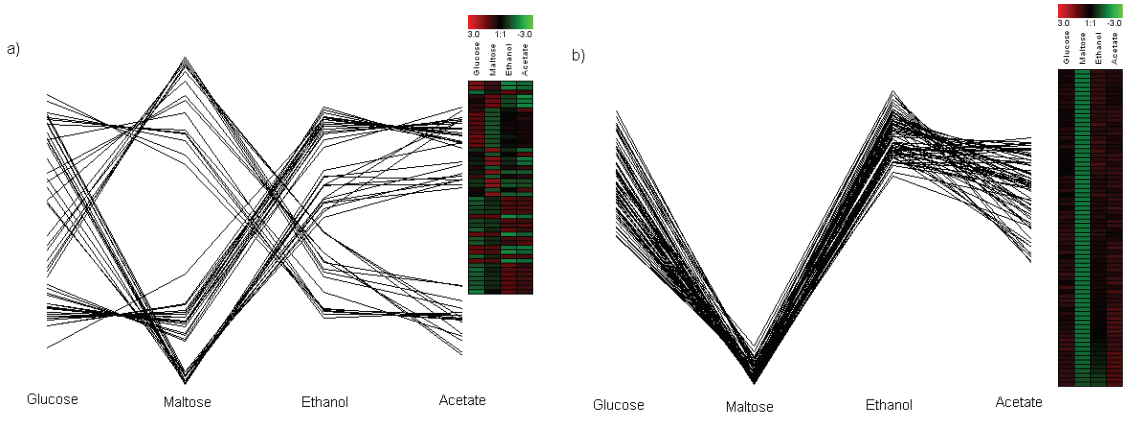


Figure S-3.3. Gene expression profiles for (a) one of the high scoring subnetworks and (b) one of the high scoring clusters obtained with SOM.

3.8.4 Robustness of the algorithm towards removal of co-factors

Redox co-factors (NAD+/NADH, NADP+/NADPH) and energy co-factors (ATP/ADP) connect different metabolic pathways to create a densely connected metabolic network. For example, the reaction interaction graph for *S. cerevisiae* has relatively high number of edges (>57000) as opposed to number of nodes (1175). Any node in this graph is on average less than 3 nodes away from any other node. From biological point of view, a perturbation at any node in the network can propagate rapidly to virtually every "corner" of the metabolic network. This demands a truly global regulatory program to act against such perturbations, as reflected in our findings on subnetworks. We also found that such regulatory actions of the cell are centred on the perturbed node, which would help to control the effects of the perturbation most effectively.

Supplementary figure S-3.4 shows the distribution of number of metabolites contributing for an edge in an enzyme interaction graph of *S. cerevisiae*. Most of the edges are contributed by 2 or 4 metabolites. Thus removal of a certain metabolite can still keep the high connectivity between nodes through other metabolites. This leads to high robustness of the subnetwork search algorithm, as network connectivity remains high even though certain edges are removed from the graph. To illustrate this, we removed certain highly connected co-factors from the metabolic network and searched for high scoring subnetwork using *GDH1* dataset. The results are summarized into Supplementary table S-3.5 that lists number of edges removed; graph diameter and overlap with the original subnetwork. The graph diameter (average distance between any two nodes) was relatively insensitive to the removal of many co-factors. Consequently, it was possible to obtain almost 75 % overlap with the original subnetwork even after removal of both redox co-factors and ATP/ADP pair, that resulted into 27 % reduction in the number of edges. The most sensitive deletion was that of NADP+/NADPH, which is consistent with the fact that GDH1 encodes for NADPH dependent enzyme. Remarkably, the deletion of NAD+/NADH was less influential, even though it resulted into removal of more edges from the network. This once again illustrates the principle of "changes around the perturbed node" and effectiveness of our algorithm to capture these changes. It should also be noted that the *reporter metabolite* analysis is insensitive to metabolite removal, except for the metabolites that are removed.

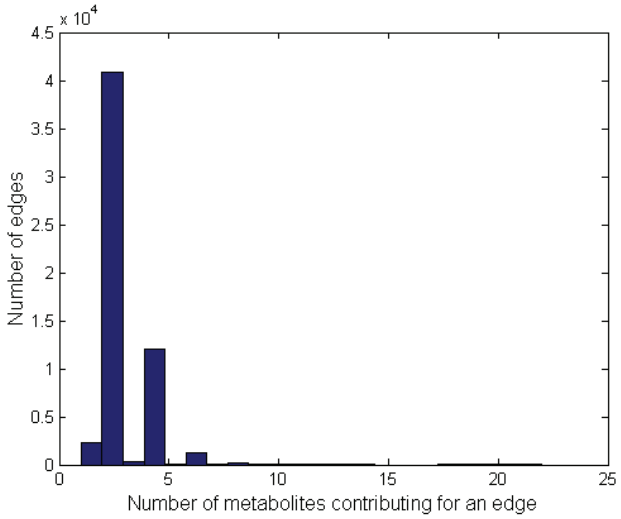


Figure S-3.4. Number of metabolites contributing for an edge in an enzyme interaction graph of *Saccharomyces cerevisiae*.

3.8.5 Distribution of *Reporter metabolites* obtained from several datasets

We identified the r*eporter metabolites* for around 47 transcriptional datasets for the yeast *S. cerevisiae.* Supplementary table S-3.6 lists the corresponding *reporter metabolites*. The average rank of any metabolite (defined as arithmetic average of ranks of a metabolite, based on $Z_{metabolite}$ score, from all conditions analysed) was more than 150 (Supplementary figure S-3.5 shows the complete distribution) illustrating the uniqueness of the *Reporter metabolites* for the particular experiment. Although we found that certain metabolites appeared frequently as *reporter*, about one-third of the metabolites were *reporter* in at least one of the datasets (Supplementary figure S-3.6 shows the complete distribution).

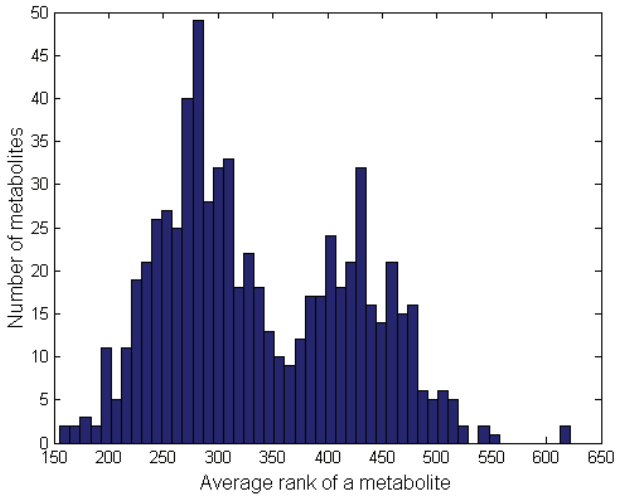


Figure S-3.5. Histogram of the average rank of a metabolite in several different datasets analysed.

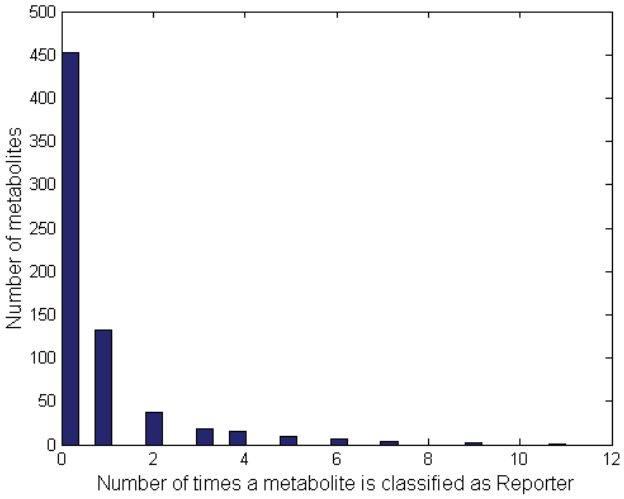


Figure S-3.6. Histogram of the frequency of a metabolite being classified as a *reporter metabolite* in several different datasets analysed.

3.8.6 Supplementary note

A key requirement for application of our method is the availability of a genome-scale model for the metabolism, but the metabolic networks have been reconstructed for several microorganisms (Price et al., 2003) and this opens for a wide application of our method for analysing different cellular systems. If the metabolic network has not been reconstructed for a given cellular system one may rapidly obtain at least a rough network structure from KEGG (Kanehisa et al., 2004) which would still result into good subnetwork analysis due to the robustness of the algorithm towards the missing information.

Table S-3.1. *Reporter metabolites* for pair-wise comparison of the different carbon sources.

Glucose-Maltose			**Glucose-Ethanol**		
alpha-D-Glucose	28	6	beta-D-Fructose 6-phosphate	15	5
FRUxt	12	-	GLCxt	14	-
Maltose	4	1	alpha-D-Glucose	28	6
MANxt	11	-	6-Phospho-D-gluconate	5	1
Ethanol	5	1	FumarateM	4	-
GLCxt	14	-	alpha-D-Mannose	15	2
Oxygen	16	2	FRUxt	12	-

3-Phospho-D-glyceroyl phosphate	5	2	2-Oxoglutarate	21	12
UbiquinolM	9	-	Carnitine	3	1
Methanethiol	1	-	Urea	4	4
Glucose-Acetate			**Maltose-Ethanol**		
Isocitrate	5	3	Carnitine	3	1
alpha-D-Glucose 6-phosphate	11	6	Allantoate	3	1
beta-D-Fructose 6-phosphate	15	5	dUTP	3	1
alpha-D-Glucose	28	6	Maltose	4	1
6-Phospho-D-gluconate	5	1	CARxt	1	-
Chitosan	2	1	Glyoxylate	6	5
alpha-D-Mannose	15	2	Oxaloacetate	8	9
Glyoxylate	6	5	dCTP	2	1
D-Glucose 1-phosphate	7	7	Ergosta-5,7,24(28)-trienol	2	1
beta-D-Fructose 1,6-bisphosphate	4	4	ITP	2	1
Maltose-Acetate			**Ethanol-Acetate**		
Glyoxylate	6	5	Chitosan	2	1
Ergosta-5,7,24(28)-trienol	2	1	alpha,alpha-Trehalose	4	1
Acetyl-CoA	19	23	2-Dehydro-3-deoxy-D-arabino-heptonate 7-phosphate	3	1
3-Phospho-D-glyceroyl phosphate	5	2	(S)-LactaldehydeM	1	-
Chitosan	2	1	N-(L-Arginino)succinate	2	3
Maltose	4	1	Chitin	5	1
Malate	8	7	NADHM	12	-
1L-myo-Inositol 1-phosphate	2	3	Carbamoyl phosphate	3	6
CO2M	12	-	AcetateM	2	-
CarnitineM	2	-	dUMP	4	1

Only the top ten scoring metabolites are shown. The first number behind the *reporter metabolite* is the number of neighbors to the *reporter metabolite* (or the number of reactions in which the reporter metabolite participates in) and the second number is the number of KEGG pathways in which the *reporter metabolite* appearsThe metabolite names ending with 'M' and 'xt' indicate that the metabolite is present in mitochondrial compartment and extracellular medium respectively. Since KEGG pathways do not classify metabolites in this fashion, the corresponding fields in the table are empty.

Table S-3.2. Genes included in the larger subnetworks obtained by analysis of gene expression datasets for *gdh1Δ*, *grr1Δ* and carbon-sources.

gdh1Δ	***grr1Δ***	**Carbon source**
HXK1 PMP2 VHT1 AGP3 PHO84 TDH1 AAC3 ORT1 CAR1 MSR1 PET9 DIC1 SFC1 YEL047C FLX1 YML082W MEP3 DUR1 MEP2 DAL3 MEP1 URA8 HEM3 GDH3 YEL041W IDH1 PUT1 UTR1 KGD1 ALD4 MET13 MAE1 MET12 TRR2 ARG5 ARG8 AAT1 LYS9 ADE3 QNS1 SRT1 UGP1 RAM1 IPP1 NPT1 MSK1 LSC1 LEU4 ILV3 CDS1 HTS1 HIP1 HNM1 CKI1 MET3 HXK2 HXT5 SUC2 GAL2 HXT14 HXT1 MAL32 EXG1 FKS1 BGL2 EXG2 GAL10 TPS3 CHS2 TPS1 YNK1 FOL2 YJL068C GPX2 URA1 HEM14 HEM15 SCS7 BNA1 CTT1 YJR078W VAP1 UGA4 HEM2 ITR1 LYP1 PUT4 BAP2 SAM1 MET1 ERG6 BIO3 SAH1 YJR105W THR1 PFK2 DED81 ASP1 GNP1 LCB1 PPT2 CIT2 HMG2 ALD6 IDP3 BAT2 GFA1 SER1 GDH1 CYS3 YIL167W TRP5 TKL1 ARO3 RIB5 FBP1 RHR2 GPD2 ADH5 GPD1 FDH1 HOM6 ZWF1 TRR1 ARA1 GND1 ERG4 NCP1 ERG9 TSC10 LCB4 ASN2 PFK1 APA2 HOR2 INM1 HIS2 PFK26 CYR1 DUT1 PCT1 ERG20 MUQ1 CPT1 PRS2 PRS1 PGM1 TSL1 PGI1 YGR043C ERG26 PDC5 PSD2 SOL2 SOL3 FUN63 YAR075W ADE17 SHM2 IMD3 PUR5 ENO2 ENO1 PCM1 DAL7 MLS1 YLR231C DYS1 NDI1 ADH3 SPE4 URA4 DPM1 PMT2 ILV2 PSD1 GCV2 ACO1 DAL2 DAL1 DAL4	*GLK1 PMA1 THI7 DIP5 FEN2 YIL145C AAC3 YER053C TDH1 PET9 HXK1 HXT7 SUC2 HXT4 HXT6 HXK2 HXT5 MAL32 HXT3 YGR287C HXT2 NTH2 HXT16 STL1 HXT1 ADE4 COX10 PSA1 YGL245W AAC1 PFK2 QNS1 IDH1 ODC1 IDP2 ADE3 PFK1 FOL3 PGK1 THI20 CDC19 TRP3 ABZ1 GNP1 BPH1 ACS2 RIB1 ADE12 RHR2 GUT2 SDH3 SFC1 LSC2 ECM40 CIT1 CTP1 MAL31 YGL186C APT2 TRP4 PRS5 HIS1 PCK1 ARO3 ARO2 ARO7 FBP1 ARO4 ENO2 PYC1 YDR341C FAB1 ARG1 MVD1 ARG3 CAR1 FBP26 CPA2 ARO8 BAP3 FCY2 TAT2 BIO5 BIO2 HIP1 MUP1 MHT1 SPE3 DYS1 NDI1 ALD4 MAE1 PUT2 CYB2 RIP1 URA1 GPX1 HYR1 BNA1 SUR2 ALD6 GND1 LYS9 ARO9 LYS21 NAT1 MLS1 YFL030W YFR055W PNC1 ILV1 GDH2 HIS7 ADE6 KRS1 ADE1 MET22 TPS2 SAM1 GLC3 THR4 SER1 LYS1 MDH3 LEU2 PDC5 GAD1 ASN2 THR1 GUA1 DED81 ASP1 YMR293C GDH3 HOM2 GPH1 SER2 SHM2 ADE8 URA2 ASN1 URA5 DUT1 YNK1 KRE2 RNR1 GSC2 TPS1 CHS2 TSL1 URA6 XKS1 TKL2 PUS1 PGM2 INO1 PGI1 PMI40 FUR1 YJR105W GLN1 AAT2 LYS20 CAT2 CRC1 ACH1 ERG10 KGD1 MDH1 FUM1 OSM1 LSC1 YCR024C POT1 FOX2 SER33 ADH4 SER3 ADH5 GLT1 ADH1 YAR075W NDH1 IMD4 FUN63 PUR5 GSY1 GSY2 TRR1 ECM17 ERG3 ERG24 ERG25 ERG2 URA4 LEU1 MET6 HIS3 ICL1 SOL1 SOL4 SOL3 ACO1 YJL200C*	*GLK1 PMA2 HXK1 PMA1 PFK1 PRS5 FOL1 PCK1 GCV1 PNC1 YMR293C FMT1 DFR1 GND2 HOM2 PHO84 TDH1 GPD2 MDH2 DIC11 SFC1 FUM1 MLS1 YFL030W ICL1 PDC6 IDP3 GDH2 DIP5 GNP1 GFA1 HIS7 UGA1 HOM6 ERG24 NDH1 COQ3 COQ6 GND1 ERG5 PAN5 NCP1 YBR006W PRO2 ECM17 ERG9 ERG3 BNA1 YJR078W ERG11 CTA1 HYR1 ERG25 HMG1 NAT2 LYS20 SER1 AGP1 FEN2 MAL31 NHA1 HNM1 BPH1 SAM3 UGA4 ITR1 TAT1 AAT2 BAT2 BAP2 MHT1 SPE3 OPI3 MET6 CYS4 BIO2 CIT2 CAT2 AGP2 CRC1 ACS1 RAM1 COX10 MET22 ARO2 ARO7 TPS2 THR4 ADE1 YER053C LSC1 MSE1 PPA2 MSW1 INM1 GSH1 FAD1 FLX1 POX1 GUT2 FBA1 TPI1 GUT1 FAA2 PRO1 TOR1 HTS1 PIK1 YNK1 SEC59 PMT5 QRI1 CYR1 APT2 PRS2 FUR1 URA5 RNR1 KTR2 DPM1 PMT1 FKS1 FUN63 PDE2 YJR105W SAM1 ARO1 URK1 KRE2 FBP26 HIS2 RIB5 ADE6 MET14 DED81 THR1 HOM3 HIS4 ERG8 ACH1 GAL7 PGM1 RKI1 TKL2 TAL1 PIS1 IPT1 CHO2 YLR089C ARG8 IDH1 PSD1 ADH3 ALD5 PUT2 ACO1 DLD1 HMGS ERG2 ADH1 FDH1 NDH2 ECM31 SOL3 ATP1 ENO2 GPM3 HXT7 SUC2 HXT3 NTH2 HXT5 HXT2 HXT10 HXT8 MAL32 FSP2 YGR287C YJL216C*

Table S-3.3. Genes included in the secondary subnetwork for pairwise comparison of the different carbon sources.

Glucose-Maltose	Glucose-Ethanol	Glucose-Acetate
FBA1 TDH2 CPA2 HEM13 PDC1 SPE1 ADH1 YAR075W SFA1 IMD3 ADH2 PUR5 NDH1 COQ3 RIP1 DYS1 ACP1 HEM1 CIT1 LEU4 ALD4 POX1 YNK1 MUQ1 URA5 ACS2 MET17 GNP1 FCY2 MAL31 MMP1 PDE1 ADE1 PFK26 DPL1 GSY1 ADE17 ERG3 BNA1 TPI1 MAL32 HXT8 FSP2 HXT1 HXT4	*HXK1 PMA2 TDH1 ALD3 CDA2 MDH2 FDH1 GCV1 MEP1 GDH3 GDH1 STL1 HXT5 YJL216C HXT2 HXT4 UGA1 IDP2 YPL276W PCK1 GND1 GND2 CIT2 CAT2 SFA1 NDH1 DYS1 ADH1 GCV2 IDH1 KGD1 YAT1 CIT1 FUM1 SFC1 TAL1 FBP1 TKL2 BAP2 DUR3 DAL2 DAL5 SAM3 BIO5*	*HXK1 HXT7 YJL216C HXT6 NTH1 HXT2 CPA2 PFK2 GLN1 TDH1 FDH1 IDP2 GDH3 GND1 ERG1 HEM14 YPL276W MDH2 NDH2 SDH3 FUM1 MLS1 LYS20 YAT1 CRC1 IDH1 PSD1 ACO1 CIT2 ACH1 CDA2 ACS1 CDA1 TRP4 PCK1 PFK26 FBP1 TAL1 ICL1 ICL2 ADH1 RKI1 SOL4 INO1*
Maltose-Ethanol	**Maltose-Acetate**	**Ethanol-Acetate**
TDH1 PHO84 PHO11 MAL31 DUR3 DAL2 DAL5 MUP3 SAM3 FUM1 SFC1 MDH2 YEL041W PDA1 GCV2 YAT1 CRC1 AGP2 YCR024C MET12 PCK1 AAT2 STL1 HXT8 MAL32 FSP2 GDH3 HEM3 DCD1 YNK1 DUT1 ACS1 CDA2 ADK1 CIT2 ERG5 FDH1 PDC1 YPL276W ADH1	*GLK1 HXT8 MAL32 FSP2 CPA2 PHO11 TDH1 SAM1 PFK2 POX1 DUT1 DCD1 PDX3 ERG3 GDH3 GND1 ERG5 FDH1 TDH2 SER2 FBP1 TPS2 YPL276W MDH2 MAL31 FUM1 MLS1 ACH1 CDA2 CIT2 PCK1 CDC19 ICL1 ICL2 NDH2 ACP1 PSD1 IDH1 ACO1 ADH1 RKI1 OPI3 INO1*	*CDC19 AAC1 TPS2 GSH2 GPX1 PDX3 MEP1 ASP1 ARG1 ARG4 FUM1 MDH1 ALD5 PDA1 IDH1 ACO1 ARO1 CKI1 CPA2 ARO3 HOM2 SAM1 GLT1 ADH1 PDE2 ATH1 TPS3 CHS3 CDA2 CDA1*

Table S-3.4. Comparison of cluster analysis of carbon source data with subnetwork obtained using here reported algorithm. Number of genes common between high scoring subnetwork and highest scoring clusters obtained with different clustering methods are listed.

Clustering method	Size of cluster	Number of genes common with subnetwork
K-means (k=10)	144	13
K-means (k=15)	83	13
K-means (k=20)	66	0
SOM (9 nodes)	272	31
SOM (25 nodes)	134	13
SOM (49 nodes)	72	8

Table S-3.5. Effect of metabolite removal on the connectivity in the reaction-interaction graph and subnetwork search.

Metabolites removed	Number of edges	Diameter of reaction interaction graph	% Overlap with the high scoring subnetwork for gdh1 case
None	57217	2.7289	100 (34/34)
NAD+, NADH	55686 (97.3 %)	2.7584	91 (31/34)
NADP+, NADPH	55601 (97.1 %)	2.7583	85 (29/34)
ATP, ADP	44733 (78 %)	2.9357	88 (30/34)
NAD+, NADH, NADP+, NADPH	54059 (94.4 %)	2.7773	79 (27/34)
NAD+, NADH, NADP+, NADPH, ATP, ADP	41554 (72.6 %)	3	73 (25/34)

Table S-3.6. *Reporter metabolites* for different gene expression datasets. (Please see Supplementary table S-3.7 for the short description and the source of the data)

Data-Title	Reporter metabolites
AER-C-N	Glyoxylate, L-Asparagine, L-Alanine, SERxt, ASNxt, GLNxt, ALAxt, GLUxt, L-Proline, GLYxt
AER-C-P	Glyoxylate, Isocitrate, sn-Glycerol 3-phosphate, FADH2M, Guanosine, S-Methylmethionine, MMETxt, L-Phenylalanine, METxt, Malonyl-[acyl-carrier protein]
AER-C-S	NADH, NAD+, GTP, Sulfate, Adenylylsulfate, O-Acetyl-L-homoserine, D-Ribose 1-phosphate, L-Homoserine, 2-Amino-3-carboxymuconate semialdehyde, 2,3-Dehydroacyl-[acyl-carrier-protein]
AER-N-P	ALAxt, GLYxt, L-Tyrosine, L-Leucine, L-Phenylalanine, HISxt, L-Isoleucine, L-Valine, METxt, PROxt
AER-N-S	H+EXT, NH3xt, L-Asparagine, L-Aspartate, 2-Hydroxybutane-1,2,4-tricarboxylate, L-Cysteine, CYSxt, GLYxt, 2-OxoglutarateM, Sulfate
AER-P-S	Pyruvate, Glycogen, Sulfate, Sulfite, alpha,alpha'-Trehalose 6-phosphate, UDP, METxt, Adenylylsulfate, 1-(5'-Phosphoribosyl)-5-amino-4-(N-succinocarboxamide)-imidazole, N6-(1,2-Dicarboxyethyl)-AMP
ANA-C-AER-C	NAD+, GLUxt, NADH, Isocitrate, AcetaldehydeM, L-Alanine, SERxt, Ferricytochrome cM, Ferrocytochrome cM, ALAxt
ANA-C-N	IMP, L-Asparagine, ILExt, VALxt, R-S-Alanylglycine, Cys-Gly, PHExt, L-Phenylalanine, LEUxt, TYRxt

ANA-C-P	Thiamin, THMxt, L-Histidine, HISxt, sn-Glycerol 3-phosphate, D-4'-Phosphopantothenate, Xanthosine 5'-phosphate, L-Alanine, Tetrahydrofolyl-[Glu](n), L-Asparagine
ANA-C-S	Sulfate, SLFxt, 2-Amino-4-hydroxy-6-hydroxymethyl-7,8-dihydropteridine, 2-Amino-7,8-dihydro-4-hydroxy-6-(diphosphooxymethyl)pteridine, Dihydropteroate, 4-Aminobenzoate, Glycolaldehyde, AcetaldehydeM, IsocitrateM, 3'-Phosphoadenylylsulfate
ANA-N-AER-N	H+M, OrthophosphateM, ADPM, Ferrocytochrome cM, Ferricytochrome cM, ATPM, UbiquinolM, Ubiquinone-9M, L-Alanine, (R)-LactateM
ANA-N-P	METxt, L-Alanine, L-Phenylalanine, sn-Glycerol 3-phosphate, L-Asparagine, Glyoxylate, PHExt, ASNxt, GLNxt, ALAxt
ANA-N-S	Glyoxylate, METxt, H+EXT, PHExt, NADPHM, L-Alanine, ALAxt, MALxt, SLFxt, GLYxt
ANA-P-AER-P	Oxygen, UbiquinolM, Ubiquinone-9M, H+M, Ferricytochrome cM, Ferrocytochrome cM, 6-Phospho-D-gluconate, OrthophosphateM, CO2M, Oxaloacetate
ANA-P-S	Orthophosphate, L-Cystathionine, Sulfite, Sulfate, 3-Phosphonooxypyruvate, Hydrogen sulfide, 3-Hydroxyanthranilate, Adenylylsulfate, NADP+, 3-Phospho-D-glycerate
ANA-S-AER-S	OrthophosphateM, ADPM, H+M, NADP+M, Ferrocytochrome cM, Ferricytochrome cM, ATPM, NADPHM, IsocitrateM, NADP+
GDS104	2-Dehydro-3-deoxy-D-arabino-heptonate 7-phosphate, OxaloglutarateM, 6-Phospho-D-gluconate, dTDP, Oxalosuccinate, D-Ribulose 5-phosphate, 2,5-Diamino-6-hydroxy-4-(5'-phosphoribosylamino)-pyrimidine, D-erythro-1-(Imidazol-4-yl)glycerol 3-phosphate, GTP, Oxaloglutarate
GDS108	1-Phosphatidyl-1D-myo-inositol 4-phosphate, 10-Formyltetrahydrofolate, 1-(5'-Phosphoribosyl)-5-amino-4-(N-succinocarboxamide)-imidazole, IsocitrateM, 1-Phosphatidyl-D-myo-inositol 4,5-bisphosphate, Glutathione, 1-(5'-Phosphoribosyl)-5-amino-4-imidazolecarboxamide, CMP, 5,10-Methenyltetrahydrofolate, Guanine
GDS109	Xanthosine 5'-phosphate, (R)-5-Phosphomevalonate, Glutathione, D-Ribose 1-phosphate, H2O2, alpha-D-Mannose 1-phosphate, (R)-S-Lactoylglutathione, alpha-D-Glucose 6-phosphate, NADH, Ergosta-5,7,24(28)-trienol
GDS113	5-Phospho-alpha-D-ribose 1-diphosphate, GMP, D-Ribose 5-phosphate, 5,10-MethylenetetrahydrofolateM, alpha,alpha'-Trehalose 6-phosphate, beta-D-Fructose 1,6-bisphosphate, AMP, UDPglucose, L-Glutamine, alpha-D-Glucose 6-phosphate
GDS114	PyrophosphateM, NADP+, 3-Phospho-D-glyceroyl phosphate, NADPH, AMPM, Maltose, ATPM, Pyrophosphate, Xanthosine 5'-phosphate, Dolichyl phosphate
GDS115	5-Phospho-alpha-D-ribose 1-diphosphate, D-Glyceraldehyde 3-phosphate, PhosphatidylserineM, Xanthosine 5'-phosphate, Carnitine, IMP, 3-Phospho-D-glyceroyl phosphate, D-Ribose 5-phosphate, alpha,alpha-Trehalose, Adenylylsulfate
GDS124	Xanthosine 5'-phosphate, D-Galactose 1-phosphate, Ergosta-5,7,24(28)-trienol, Cytosine, Ubiquinone-9M, CMP, Episterol, 2-Phospho-D-glycerate, CYSxt, Inosine
GDS16	NADPHM, Ubiquinone-9M, UbiquinolM, alpha,alpha-Trehalose, NADHM, D-Glucono-1,5-lactone 6-phosphate, Carnitine, 3-Phospho-D-glyceroyl phosphate, 2-Hydroxybutane-1,2,4-tricarboxylate, alpha-D-Glucose 6-phosphate

GDS19	beta-D-Fructose 6-phosphate, 3-Phosphonooxypyruvate, Xanthosine 5'-phosphate, IMP, 5,10-Methylenetetrahydrofolate, THMxt, Sulfite, 5-Phosphoribosylamine, L-Asparagine, Phosphoenolpyruvate
GDS20	Xanthosine 5'-phosphate, D-Ribose 5-phosphate, AMP, Homocysteine, 3-Phospho-D-glyceroyl phosphate, ATP, Orotate, N6-(1,2-Dicarboxyethyl)-AMP, Phytosphingosine, Orotidine 5'-phosphate
GDS21	beta-D-Fructose 6-phosphate, Orthophosphate, 5-Phospho-alpha-D-ribose 1-diphosphate, Sedoheptulose 7-phosphate, D-Glyceraldehyde 3-phosphate, Glycogen, D-Erythrose 4-phosphate, IMP, 2-OxoglutarateM, Adenosine
GDS30	AMP, Xanthosine 5'-phosphate, alpha-D-Glucose, Maltose, UDP, 5-Phospho-alpha-D-ribose 1-diphosphate, IMP, THMxt, 3-Phospho-D-glyceroyl phosphate, 1,3-beta-D-Glucan
GDS32	alpha-D-Glucose, GLCxt, NAD+M, FRUxt, MANxt, Succinate, D-Galactose, myo-Inositol, NADHM, Ethanolamine phosphate
GDS33	D-Fructose, GLCxt, FRUxt, alpha-D-Mannose, MANxt, alpha-D-Glucose, Acetaldehyde, L-Threonine, 2-Oxoglutarate, Mannan
GDS354	5-Phospho-alpha-D-ribose 1-diphosphate, 6-Phospho-D-gluconate, alpha,alpha'-Trehalose 6-phosphate, L-Glutamine, Glycogen, Spermidine, AMP, 1-(5'-Phosphoribosyl)-5-amino-4-(N-succinocarboxamide)-imidazole, Zymosterol, 5-Phosphoribosylamine
GDS362	Xanthosine 5'-phosphate, Isocitrate, Acetaldehyde, Ethanol, Pyruvate, (S)-LactateM, 3',5'-Cyclic AMP, Uracil, Maltose, Tetrahydrofolyl-[Glu](n)
GDS37	D-Erythrose 4-phosphate, Sedoheptulose 7-phosphate, Xanthosine 5'-phosphate, Glutathione, D-Glyceraldehyde 3-phosphate, 4,4-Dimethylzymosterol, 4-Aminobutanoate, dADP, Acetate, beta-D-Fructose 6-phosphate
GDS457	FRUxt, alpha-D-Mannose, MANxt, D-Fructose, alpha-D-Glucose, GLCxt, H+EXT, Maltose, L-Serine, Mannan
GDS600	ATP, 1-Phosphatidyl-D-myo-inositol, Isopentenyl diphosphate, ADP, 3-Nonaprenyl-4-hydroxybenzoate, Glycogen, all-trans-Nonaprenyl diphosphate, 3-Phospho-D-glyceroyl phosphate, tRNAM, (R)-5-Phosphomevalonate
GDS608	NADP+M, ADPM, Xanthosine 5'-phosphate, D-Ribose 5-phosphate, ATPM, OrthophosphateM, NADPHM, alpha,alpha'-Trehalose 6-phosphate, 2-OxoglutarateM, 2-Phospho-D-glycerate
HAP1-WT	S-Adenosyl-4-methylthio-2-oxobutanoate, 8-Amino-7-oxononanoate, 7,8-Diaminononanoate, H+EXT, CoA, Uracil, Cytosine, Palmitoyl-CoA, tRNA(Lys), L-lysyl-tRNA(Lys)
ROX1-WT	5,10-MethenyltetrahydrofolateM, 10-FormyltetrahydrofolateM, FormateM, TetrahydrofolateM, Sodium, NAxt, L-Cysteine, ADPM, CYSxt, OrthophosphateM
GDS69	D-Glucose 1-phosphate, 2-Phospho-D-glycerate, D-Fructose, alpha-D-Glucose, TRPxt, MANxt, alpha-D-Glucose 6-phosphate, 3-Phospho-D-glyceroyl phosphate, H+EXT, GLCxt

Table S-3.7. Short description and source of the datasets used for the analysis reported in the Supplementary table S-3.6.

Data-Title	Data-type*	Short description	Source**
AER-C-N	D	Aerobic chemostat, Carbon-limited Vs. Nitrogen limited.	1
AER-C-P	D	Aerobic chemostat, Carbon-limited Vs. Phosphorous limited.	1
AER-C-S	D	Aerobic chemostat, Carbon-limited Vs. Sulphur limited.	1
AER-N-P	D	Aerobic chemostat, Nitrogen-limited Vs. Phosphorous limited.	1
AER-N-S	D	Aerobic chemostat, Nitrogen -limited Vs. Sulphur limited.	1
AER-P-S	D	Aerobic chemostat, Phosphorous -limited Vs. Sulphur limited.	1
ANA-C-AER-C	D	Chemostat, Aerobic Vs Anaerobic, Carbon-limited	1
ANA-C-N	D	Anaerobic chemostat, Carbon-limited Vs. Nitrogen limited.	1
ANA-C-P	D	Anaerobic chemostat, Carbon-limited Vs. Phosphorous limited.	1
ANA-C-S	D	Anaerobic chemostat, Carbon-limited Vs. Sulphur limited.	1
ANA-N-AER-N	D	Chemostat, Anaerobic Vs Aerobic, Nitrogen-limited	1
ANA-N-P	D	Anaerobic chemostat, Nitrogen-limited Vs. Phosphorous limited.	1
ANA-N-S	D	Anaerobic chemostat, Nitrogen -limited Vs. Sulphur limited.	1
ANA-P-AER-P	D	Chemostat, Anaerobic Vs Aerobic, Phosphorous-limited	1
ANA-P-S	D	Anaerobic chemostat, Phosphorous -limited Vs. Sulphur limited.	1
ANA-S-AER-S	D	Chemostat, Anaerobic Vs Aerobic, Sulphur-limited	1
GDS104	M	Temporal analysis of the developmental program of sporulation	2
GDS108	M	Menadione exposure time course	2
GDS109	M	Hydrogen peroxide response time course	2
GDS113	M	Dithiothrietol exposure time course	2
GDS114	M	Stationary phase time course	2
GDS115	M	Amino acid and adenine starvation time course	2

GDS124	M	Cell cycle, cdc15 block-release time course	2
GDS16	M	Heat shock from 25 C to 37 C time course	2
GDS19	M	Nitrogen depletion time course	2
GDS20	M	Hyper-osmotic shock time course	2
GDS21	M	Carbon sources (glucose, raffinose, galactose, fructose, sucrose or ethanol)	2
GDS30	M	Diamide treatment time course	2
GDS32	M	Steady-state temperature (17 C, 21 C, 25 C, 29 C and 37 C, compared to cells grown at 33 C)	2
GDS33	M	Hypo-osmotic shock time course	2
GDS354	M	Lithium response in yeast	2
GDS362	M	Aging in yeast	2
GDS37	M	Diauxic shift time course	2
GDS457	M	Response to constitutive activation of the Ras/cAMP signal transduction pathway	2
GDS600	M	Deubiquitinating enzyme UBP10 inactivation	2
GDS608	M	Filamentous-form growth on solid media	2
HAP1-WT	D	Wild-type Vs Hap1-Null	3
ROX1-WT	D	Wild-type Vs Rox1-Null	3
GDS69	M	Conditions of excess copper or copper deficiency	2

* D : Differential
M: Multi-dimensional

** 1. Boer, V. M., de Winde, J. H., Pronk, J. T. & Piper, M. D. (2003) J Biol. Chem 278, 3265-3274.
2. NCBI-GEO Datasets (http://www.ncbi.nlm.nih.gov/entrez/)
3. Ter Linde, J. J. & Steensma, H. Y. (2002) Yeast 19, 825-840.

Chapter 4: Optimality assessment and performance improvement of simulated annealing algorithm for finding biologically active subnetworks

Manuscript describing the results in this chapter is under preparation.

"Are we nearly there?" Alice managed to pant out at last. "Nearly there!" the Queen repeated. "Why, we passed it ten minutes ago! Faster!"

4.1 Abstract

Bio-molecular networks of functional and physical interactions are being increasingly used as a platform to integrate genome-scale omics datasets. One of the interesting problems in such an integrative analysis is to identify connected subnetworks in large bio-molecular interaction networks that show maximum correlated response to a perturbation. This problem is NP-hard and hence usually attempted with the stochastic algorithms such as simulated annealing. However, simulated annealing does not guarantee to find the global optimal solution. In the present work we first report a method to estimate the upper bound on the global optimal score. Using the principles underlying this method we propose a set of heuristics that significantly improved the performance of a simulated annealing algorithm. We demonstrate the applicability of the proposed methodology for analysis of gene expression data for the yeast *Saccharomyces cerevisiae*. Two important bio-molecular interaction networks, protein-protein and enzyme interaction network, were used as data integration platform.

4.2 Background

Biological systems at cellular level can be viewed as an interaction network of molecules operating towards a set of objectives. This network-view of the cell is not only a convenient visual representation but also offers a conceptual framework for the modeling of the complex cellular systems. Interactions in such networks arise not only due to the physical contact between two molecules (e.g. protein-protein, protein-DNA interaction) but also due to the functional associations (e.g. synthetic lethality, metabolic function). Bio-molecular networks form the basis of several systems biology approaches and have been successfully used to uncover the underlying transcriptional regulatory networks in the protein-protein/protein-DNA interaction networks (Ideker et al., 2002) and metabolic networks (Patil and Nielsen, 2005) (Chapter 3). Consequently gene expression data analysis has moved to a new era where integration with the biological networks is an essential part of the analysis. However, this methodology is not limited for the gene expression data alone but can also be used for the analysis and integration of several other omics datasets, such as proteomics (unpublished results) and metabolomics (Chapter 5).

One of the key problems in such an integrative data analysis is to identify connected subnetworks with maximum collective response to a perturbation, which can be quantified a subnetwork score. The score for each individual node in the network is based on its response observed in the experiment. This problem is NP-hard (Ideker et al., 2002) and usually solved with stochastic algorithms such as simulated annealing. However, simulated annealing algorithm does not guarantee

to find the global optimal solution and consequently may lead to an incomplete picture of the underlying biological phenomenon. Moreover, the "degree of optimality", or the closeness of the final solution to the global optimum is not evident from the algorithm, thus making it difficult to access the significance of the solution. In the present work we address this issue from an algorithmic point of view. We first developed a method to estimate the upper and the lower bound on the global optimal score. Apart from providing an useful measure of the "degree of optimality" for final solution, we use the underlying principles to suggest a new set of heuristics. We show that these heuristics result in significant performance improvement for simulated annealing. The applicability of the proposed methodology is illustrated for two examples from each of the two important classes of the biological networks, namely, protein-protein interaction networks and metabolic networks (enzyme interaction networks (Patil and Nielsen, 2005) (Chapter 3)).

4.3 Problem definition

The bio-molecular interaction network under study can be regarded as a graph with nodes representing molecules (e.g. proteins) and interaction between them as edges. Each node then can be assigned a score based on its response observed in a particular experiment. Node scores were calculated based on the significance of change in the gene expression levels between two conditions (or between two mutants). P-value of each node, obtained by using student's t-test (p_i), can be converted to a Z-score by using inverse normal cumulative distribution function (CDF, θ^{-1}).

$$Z_{ni} = \theta^{-1}(1 - p_i) \tag{1}$$

Any subgraph of size k can then be scored using size-independent aggregated Z-score:

$$Z_s = \frac{1}{\sqrt{k}} \sum_k Z_i \tag{2}$$

Z_s scores need to be corrected for the background distribution by subtracting the mean (μ_k) and dividing by the standard deviation (σ_k) of the aggregated Z-scores of several sets of k nodes chosen randomly from the parent graph.

$$Z_s^{corrected} = \frac{(Z_s - \mu_k)}{\sigma_k} \tag{3}$$

$Z_s^{corrected}$ characterizes the biological activity, or the aggregate transcriptional response of the subnetwork. The problem addressed here is to find the subnetwork with the maximum $Z_s^{corrected}$ score (global optimal solution).

4.3.1 Simulated annealing algorithm for subnetwork finding

Simulated annealing algorithm has been described in chapter 3 (supplementary material).

4.3.2 Interaction networks and transcriptome data used

Two biologically important cellular networks, namely protein-protein interaction (PPI) network and reaction interaction network (Patil and Nielsen, 2005) (Chapter 3), were used in this study. Both networks are from *Saccharomyces cerevisiae* (bakers yeast) which is not only an industrially important microorganism but also has been used as a model organism to study human diseases. Protein-protein interaction network used consists of 3783 nodes and 23764 edges (obtained from EMBL, http://www.embl.org). This network shows a scale-free topology (Strogatz, 2001) which is characterized by few hubs (nodes with high degree), with rest of the nodes having relatively low degree. Unlike majority of the studies, where only high confidence subset of network is considered, we used the entire network. The rationale behind this choice is that the networks of large size will become increasingly available, and the existing low-confidence interactions (several of which also represent important biological information) may find their way into search algorithms (A.P. Oliveira, personal communication). Gene expression data for a reference strain and mutant deleted in *HXK2* was used with PPI network. P-values for each node were calculated by using student's t-test.

As a second example of interaction network, an enzyme interaction graph was constructed for *Saccharomyces cerevisiae* which consisted of around 700 nodes and 50000 edges. The average number of edges per node is much higher in the reaction interaction graph as compared to the PPI graph. Since the reaction interaction graph has much lower diameter (~2.6), the number of possible subnetworks is much higher as opposed to the PPI graph with the same number of nodes. Gene expression data for a reference strain and mutant deleted in *GDH1* was used with the enzyme interaction network. P-values for each node were calculated by using student's t-test.

4.4 Results and discussion

4.4.1 Upper bound on global optimal score

We consider a hypothetical network where all nodes are connected to all other (a complete graph). Since the mean and the standard deviation of the background score (equation 3) are func-

tions of k alone (given the data), the subnetwork consisting of k top-scoring nodes will be the maximum scoring subnetwork. Consequently, the upper bound for the global optimal score can be determined by considering maximum scoring subnetworks of all sizes in the hypothetical complete graph. These results can be conveniently visualized by plotting the maximum possible score against k (hereafter referred as boundary plot). Boundary plots for the yeast PPI network (with the *HXK2* data) and the enzyme interaction network (with the *GDH1* data) is shown in figure 4.1. The region below the maximum score curve defines the "feasible" region for the subnetwork score. Thus the "degree of optimality" or the confidence on the optimality of any subnetwork solution can be estimated based on the distance from the global optimal solution and the optimal solution for size k of that subnetwork.

It is also possible to define the lower bound for the subnetwork score in the similar way as for the upper bound. Thus, the lowest scoring subnetwork of size k will be comprised of the lowest scoring k nodes. Although the lowest scoring subnetworks have not been a subject of biological importance, it may reveal some interesting features of the data. However, in this study focus is on maximum scoring subnetwork and for this purpose we define an "algorithmic lower bound" for the optimal score.

4.4.2 Algorithmic lower bound

The global optimal solution for the complete graph (the point with the maximum score on the upper bound curve) implies that if the k nodes at that point (k*) would have been forming a completely connected subnetwork in the network under question, that network would have been global optimal as well, irrespective of the topology of the rest of the network. Thus if we calculate the scores of the subnetwork (/s) formed by k* nodes, then the maximum amongst those scores is defined as the algorithmic lower bound. From the definition of upper bound curve and algorithmic lower bound it can be expected that any algorithm used should perform at least to obtain the score equal to the algorithmic lower bound.

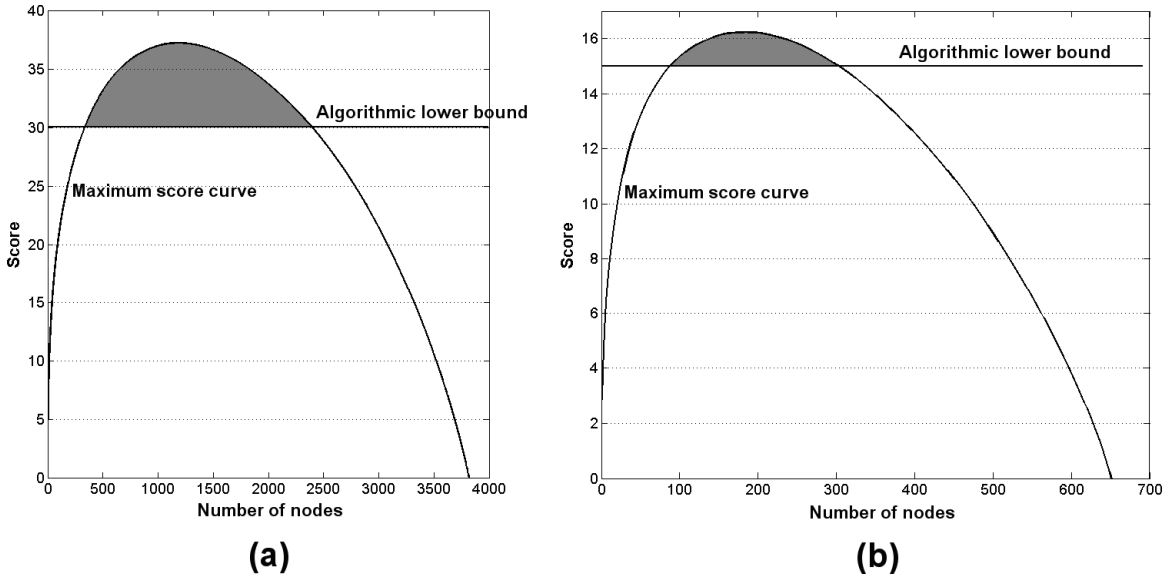


Figure 4.1. Maximum score curve and algorithmic lower bound for: (a) PPI network and HXK2 data and, (b) Enzyme interaction network and GDH1 data. Area between maximum score curve and algorithmic minimum score line (shaded area) denotes the expected performance area for subnetwork finding algorithms.

From figure 4.1 it can be seen that in case of the enzyme interaction network the area between maximum score curve and algorithmic minimum score line (shaded area) is relatively small. This is consequence of very high connectivity (per node) in the network. Moreover, the simulated annealing algorithm performed in the shaded area for enzyme interaction graph. In contrast, simulated annealing search in the PPI graph with random initial condition resulted in several solutions below the algorithmic lower bound (data not shown), showing the far from optimal nature of the solutions obtained. Hence only the PPI network was investigated further for improvement.

4.4.3 Proposed heuristics

Based on the nature of the global optimum in a complete graph, we propose the following set of heuristics to improve the performance of the simulated annealing algorithm.

Lower bound initialization

All k* nodes (forming global optimal subgraph for the given graph) are initiated to 1 (visible) and rest to 0 (invisible).

P0 initialization

Each node is initialized to 1 based on its p-value (or Z-score). Higher the Z-score, higher is the probability to be set to 1.

P1 initialization

Reporter score (similar to reporter metabolite score (Patil et al., 2004) (Chapter 2)) is calculated for each node. Reporter score is the normalized and background corrected collective score for the nearest neighbors of the protein under consideration. Then each node is initialized to 1 based on its reporter score, higher the score higher the probability to be set to 1.

P0P1 initialization

Combination of P0 and P1 initialization. A node is probabilistically set to 1 if either the Z-score or the reporter score is high.

The lower bound initialization heuristic is directly implied by the optimality curve. The other heuristics (P0, P1 and P0P1) were derived based on the stochastic nature of the simulated annealing. Since the network under consideration is not complete, the lower bound initialization may possibly lead to the local optimum. P0 heuristic may help in such cases. P1 heuristic follows from the fact that the reporter score of each node reflects how good are the neighbours of that node and hence whether it is acting as a bridge between two (or more) high scoring nodes.

4.4.4 Performance of the algorithm employing proposed heuristics

To test the proposed heuristics, we analyzed the gene expression data from HXK2 experiment for subnetwork finding in PPI graph. The results were compared with simulated annealing without employing any of the proposed heuristics (i.e. randomly initiated). In order to enable a fare comparison, all instances (including the base algorithm) were optimized for the starting temperature and the hub diameter (data not shown). The parameter optimization was also based on the results of 10 independent runs. Each instance of the algorithm (employing one of the heuristics) was run for 100 times and the resulting solutions were used to evaluate the performance. The results of the comparison are summarized in figure 4.2 which shows the mean, standard deviation and maximum scores obtained for each instance of the algorithm. Figure 4.1 (a) shows the optimality curves for the same problem.

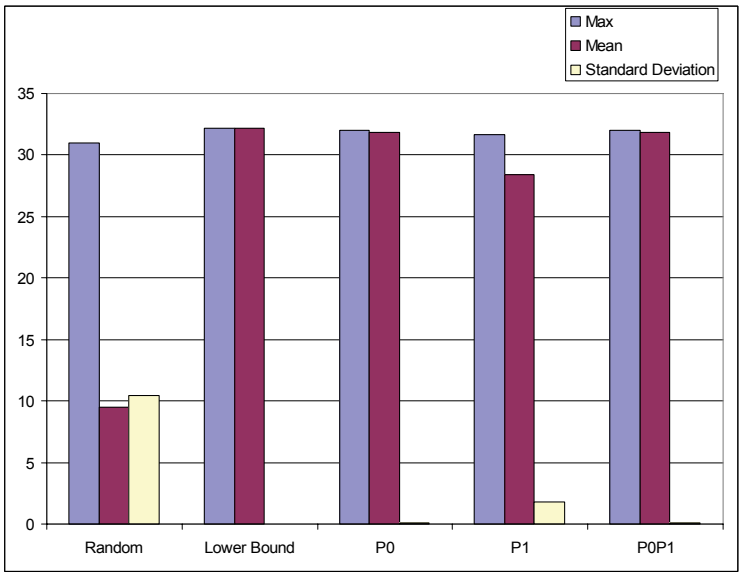


Figure 4.2. Comparison of different initialization heuristics. All heuristics performed better than random initiation. Lower bound initialization performed best.

All heuristics performed well compared to random initialization. Although the maximum score obtained from several runs is not so low for random case, very high standard deviation means that the algorithm must be run many times to get a good solution. On the other hand, very low standard deviation for heuristics means that a good solution can be obtained in first (or very few) run (/s). Moreover, the heuristics were found to be much less sensitive to simulated annealing parameters compared to random case (data not shown). Thus while implementing the heuristics; it is not necessary to do complete parameter optimization.

4.5 Conclusions

Heuristics proposed in this study seem to work very well on the chosen dataset and network. This will result in significant reduction in computational time for solving data integration problems and will also increase the confidence in the obtained solutions. However, it is necessary to test the heuristics on different networks (with different topologies) and different datasets (different distribution of p-values). Since, our heuristics are based on some theoretical considerations; it is likely that they will also perform well in these cases. The optimality curve, on the other hand, is also an important result. This will certainly help in evaluating the performance for different algorithms, and will also help in giving confidence to the results obtained. This will especially be important when strong biological conclusions are drawn based on such analysis.

Chapter 5: Integration of metabolome data with metabolic networks reveals reporter reactions

This chapter is based on the publication:
Cakir T.*, Patil, K.R.*, Önsan , Z. İ., Ülgen, K. Ö., Kırdar, B. & Nielsen, J. "Integration of Metabolome data with metabolic networks reveals reporter reactions", Molecular Systems Biology, accepted (2006). * These authors contributed equally to this work.

'What is the use of a book', thought Alice, 'without pictures or conversations?'

5.1 Extended Synopsis

Cellular metabolism, as reflected in the metabolite levels and fluxes, is an integrated result of mass balance constraints and regulation at several different levels. Consequently analysis of cellular metabolite levels, generally referred to as metabolomics, is an important step in the post-genomic era towards understanding of the biological logic behind large-scale organization and operation of cellular metabolism. Although it is now possible to quantitatively measure many intra-cellular metabolites, interpreting such data is a difficult task owing to the high connectivity in the metabolic network and inherent inter-dependency between enzymatic regulation, metabolite levels and fluxes. Here we present a hypothesis-driven algorithm (Figure 5.1) for the integration of metabolome data with topology of genome-scale metabolic models and thereby identify the reactions (reporter reactions) significantly responding to the environmental/genetic perturbations through changes in metabolite levels. The algorithm is analogous to the algorithm developed by us earlier for identification of reporter metabolites using transcriptome data (Patil and Nielsen, 2005). For demonstration of the algorithm we use two recently collected metabolome datasets for the yeast *Saccharomyces cerevisiae*, corresponding to an environmental and a genetic perturbation (Boas-Villas et al., 2005; Devantier et al., 2005), to illustrate the applicability of the algorithm.

Analytical methods available to date for metabolome measurements cover only a small fraction of the metabolites present in genome-scale metabolic models. Consequently, lack of quantitative data for several metabolites presents a major hurdle in integration of metabolome data and network topology. We therefore apply a pathway analysis based pre-processing of the genome-scale yeast model to derive a reduced model with increased fraction of measured metabolites. In this way, the yeast genome-scale model including three compartments (mitochondria, cytosol and extra-cellular space) with 844 metabolites and 1175 reactions was reduced to a two-compartment model (intra-cellular and extra-cellular space) with 178 metabolites participating in 139 reactions, which corresponded to more than 47% of the quantitative metabolome data used in this study (84 metabolites). The first dataset (Villas-Boas et al., 2005) allowed the examination of the effect of a perturbation related to an altered redox metabolism resulting from a gene deletion and aerobic/anaerobic growth; while the second dataset (Devantier et al., 2005) was used for studying the effect of very-high-gravity fermentation media on metabolic phenotype.

Significance of change for each of the measured metabolites was quantified as p-values calculated by using the u-test. To address the problems arising due to unavailability of data for over 50% of the metabolites (after pre-processing), p-values estimated from uncharacterized peaks in GC-MS spectra were randomly assigned to the 94 metabolites that remained unmeasured in the reduced metabolic model. These p-values were then converted to Z-scores which will be normally distributed for a random dataset. Each reaction in the model was then scored by using the Z-scores of its neighboring metabolites.

$$Z_{reaction} = \frac{1}{\sqrt{k}} \sum Z_{metabolite,k}$$

To account for the random assignment of scores to unmeasured metabolites, calculations were repeated 1000 times and the resultant scores were averaged. Thus, the final scores represent the significance of reactions partially independent on the true levels of the un-quantified metabolites. Top scoring reactions are hereby termed reporter reactions.

As metabolite levels are governed by changes in fluxes and enzyme activities, reporter reactions indicate the significance of how those reactions respond to the perturbation under study. Reporter scores of reactions participating in selected pathway structures across the analyzed perturbations are consistent with the previously reported findings and/or the expectations based on the type of the perturbation. The here reported algorithm thus enabled identification of key reactions in the yeast metabolism affected by genetic and environmental perturbations. Reporter reaction analysis is an attempt to infer the differential reaction significance based on metabolite measurements, and hence provides a basis for understanding the underlying cellular processes responding to the perturbations.

We also show that our method, in combination with transcriptome data (Devantier et al., 2005) may provide information on whether a given reaction is likely to be regulated at the metabolic level or at the hierarchical level (ter Kuile and Westerhoff, 2001). The Z-score of a reaction calculated by our approach can be treated as an indicator of metabolic regulation; whereas the degree of hierarchical regulation of reactions can be approximated by the Z-scores calculated based on the changes in gene expression levels. By comparing the Z-scores emerging from different omics approaches, in this case metabolomics and transcriptomics, the underlying reasoning for regula-

tion of reactions included in our reduced model could be hypothesized. For the 121 reactions in the model having corresponding genes associated with them, the analysis allowed determination of the reactions with potential regulation at metabolic, hierarchical, or at both levels. Our results indicate that although there are many metabolically regulated reactions in the network, regulation is predominantly hierarchical.

This study can be regarded as one of the first steps towards the integration of different types of omics data by using metabolic networks as a scaffold in order to understand the architecture of metabolic regulatory circuits. Furthermore, our model driven analysis forms a platform for the integration of other types of omics data, such as proteomics, and hence allow genome-scale identification of regulation in the metabolism.

REFERENCES

Devantier R, Scheithauer B, Villas-Bôas SG, Pedersen S, Olsson, L (2005) Metabolite profiling for analysis of yeast stress response during very high gravity ethanol fermentations. *Biotechnol. Bioeng.* **90:** 703-714.

Patil KR, Nielsen J (2005) Uncovering transcriptional regulation of metabolism by using metabolic network topology. *Proc Nat Acad Sci USA* **102:** 2685-2689

ter Kuile BH, Westerhoff HV (2001) Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. *FEBS Letters* **500:** 169-171.

Villas-Bôas SG, Moxley JF, Åkesson M, Stephanopoulos G, Nielsen J (2005) High-throughput metabolic state analysis: the missing link in integrated functional genomics. *Biochemical J.* **388:** 669-677.

5.2 Abstract

Interpreting quantitative metabolome data is a difficult task owing to the high connectivity in metabolic networks and inherent inter-dependency between enzymatic regulation, metabolite levels and fluxes. Here we present a hypothesis-driven algorithm for the integration of such data with metabolic network topology. The algorithm thus enables identification of reporter reactions, which are reactions where there are significant coordinated changes in the level of surrounding metabolites following environmental/genetic perturbations. Applicability of the algorithm is demonstrated by using data from *Saccharomyces cerevisiae*. The algorithm includes preprocessing of a genome-scale yeast model such that the fraction of measured metabolites within the model is enhanced, and hereby it is possible to map significant alterations associated with a perturbation even though a small fraction of the complete metabolome is measured. By combining the results with transcriptome data we further show that it is possible to infer whether the reactions are hierarchically or metabolically regulated. Hereby the reported approach represents an attempt to map different layers of regulation within metabolic networks through combination of metabolome and transcriptome data.

5.3 Introduction

One of the goals of systems biology is to obtain overall quantitative description of cellular systems. This is currently not achievable since the number of components and interactions involved in these systems is quite large resulting in a very large parameter space. Thus, methods are required to reduce the dimensionality and particularly identify key regulatory points in the many different cellular processes. Metabolism is a good starting point to develop such analysis methods as it is studied in great detail and well annotated. Furthermore, genome-scale metabolic models have been developed for many different cellular systems (Edwards and Palsson, 2000;Forster et al., 2003a;Sheikh et al., 2005), and besides their use for simulation of cellular function (Edwards et al., 2001;Famili et al., 2003;Price et al., 2004) these models can serve as scaffolds for analysis of genome-scale biological data (Covert et al., 2004;Borodina and Nielsen, 2005). This has been demonstrated recently for analysis of transcriptome data, where the use of genome-scale metabolic models enabled identification of co-regulated sub-networks and reporter metabolites (Patil and Nielsen, 2005) (Chapter 3). Although transcriptome data provides an overview of the global regulation in the metabolism, understanding of cellular physiology is incomplete without knowledge of metabolome owing to the high connectivity in metabolic networks and inherent inter-dependency between enzymatic regulation, metabolite levels and fluxes (Nielsen, 2003). Metabolites, acting as intermediates of biochemical reactions, play a crucial role within a living cell by

connecting many different operating pathways. Metabolite levels are determined by the concentrations and the properties of the surrounding enzymes, making their levels a complex function of many cellular regulatory processes in different dimensions. Thus, the metabolome represents a snapshot of the functioning metabolism of the cell and hence provides valuable information about regulation of several different cellular processes (Villas-Boas et al., 2005b). Consequently, in recent years there has been increased focus on analysis of the metabolome (Sumner et al., 2003;Bino et al., 2004;Villas-Boas et al., 2005b). Even though traditional data analysis methods like principal component analysis, clustering analysis and chemometrics have shown to be efficient for analysis of this kind of data (Raamsdonk et al., 2001;Allen et al., 2003), there are some limitations with these methods for uncovering the underlying biological principles (Weckwerth et al., 2004). Furthermore, there are still only few example studies on the use of metabolome data to understand regulatory principles in metabolism (for an example see (Kummel et al., 2006)).

Functional analysis of cellular metabolism and in particular integration of metabolome data with other omics-data demands (semi-)quantitative measurements of key metabolites. However, a problem with metabolomics is the scarcity of targeted quantitative data, and often metabolome analysis is (at best) semi-quantitative even though there is a trend towards more quantitative analysis (Nielsen and Oliver, 2005). Although it is currently not yet possible to quantify all the metabolites in a cellular system (Goodacre et al., 2004;Fernie et al., 2004), a high-throughput GC-MS method that allows semi-quantitative identification of several metabolites in *S. cerevisiae* was recently developed (Villas-Boas et al., 2005c;Devantier et al., 2005a). In the latter studies, the levels of 52 unique metabolites (out of 584 reported unique metabolites in the genome-scale yeast model (Forster et al., 2003a)) were determined in genetically different yeast strains under different environmental conditions. Specifically, metabolites playing important roles in the central carbon metabolism and amino acid biosynthesis could be identified.

In order to understand the regulatory principles underlying the changes in metabolite levels we developed an algorithm that enables integration of such quantitative metabolome data with genome-scale models by using a graph theoretical representation of the metabolism. We demonstrate the application of this algorithm for the metabolome data reported by Villas-Bôas *et al.* (Villas-Boas et al., 2005c) and Devantier *et al.* (Devantier et al., 2005a). We use the significance of changes in the metabolite levels to identify reporter reactions around which the most significant coordinated metabolite changes are observed. Reporter reaction analysis is an attempt to infer the differential reaction significance based on metabolite measurements, and hence provides a basis for understanding the underlying cellular processes responding to the perturbations. We further

demonstrate that through combination with transcriptome data, reporter reactions may provide clues on whether regulatory control at a given reaction node is at the metabolic level or at the hierarchical level.

5.4 Results and discussion

5.4.1 Model Preprocessing

Due to the large chemical diversity of the metabolome there is currently no single analytical method that enables analysis of the complete metabolome. Even the best analytical methods reported to date for metabolome analysis therefore only cover a small fraction of the metabolites present in genome-scale metabolic models. The unavailability of data for a large number of metabolites is one of the major problems associated with mapping (and hence integration) of metabolome data on to genome-scale metabolic networks. In order to overcome this fundamental problem we pre-processed the genome-scale model of Förster et al. (2003) so as to obtain a reduced model where the fraction of experimentally measured metabolites was enriched. This processing was done by systematically eliminating unmeasured metabolites from the metabolic network. We note that the model pre-processing is dependent on the metabolome data that are available, and the pre-processing will have to be done for each case. However, following the flow-chart depicted in Supplementary Figure S-5.1 this pre-processing is relatively straight forward and can easily be done also for other metabolic networks.

The yeast genome-scale model includes three compartments (mitochondria, cytosol and external space) with 844 metabolites (559 cytosolic, 164 mitochondrial, 121 external) and 1175 reactions (Forster et al., 2003a). Within the context of this model, metabolites present in more than one compartment are treated as if they are different entities in each compartment. However, the experimental data used in this analysis (and most of the datasets available to date) can only differentiate between extracellular and intracellular space. Since metabolite levels in different cellular compartments are not available, the cytosolic/mitochondrial compartmentation of the model was removed and corresponding metabolites were represented as one, with their corresponding reactions conserved. Also, there are a number of duplicate reactions due to the presence of isoenzymes in the model, and these reactions were lumped into single reactions since metabolome data alone does not provide information that enables distinction between the operations of different isoenzymes. As a result, the 'processed' model (Uncompartmented model, UNCOMP) consists of 677 metabolites (559 internal, 118 external) with 725 reactions, including transport reactions. With this model the experimental data used here amount to about 12 % of these 677 metabolites (52 internal, 32 external).

Enzyme subsets are enzymes that always operate together in fixed flux proportions at steady state (Pfeiffer et al., 1999) (Schuster et al., 2002a); often representing enzymes in linear pathways. Accordingly, the intermediate metabolites in enzyme subsets can be assumed to be similarly affected by the perturbations. The uncompartmented model (UNCOMP) was further reduced in size by using METATOOL 4.3 (Pfeiffer et al., 1999) (Dandekar et al., 2003) and thus representing each enzyme subset as a single reaction. The resulting model (Enzyme-subset model, ENZSUB-1) consists of 563 metabolites and 590 reactions and it has about 15 % of the metabolites measured within the data used. Since the removal of the metabolites in linear pathways also led to the omission of 6 measured metabolites, the reactions containing these metabolites were restored back into the ENZSUB-1 model. To further increase the fraction of the measured metabolites, potentially inactive (or potentially low flux) reactions were removed. This was done by using Flux Balance Analysis (FBA) (Varma and Palsson, 1994) (Kauffman et al., 2003) for simulation of fluxes at specific environmental conditions used in the experiments (aerobic and anaerobic batch cultivation in glucose-limited minimal media). The ENZSUB-1 model was used to simulate the fluxes with the objective of optimum growth. Then, the maximum and the minimum flux for each reaction in the model were obtained by constraining the specific growth rate between its optimum value and 50% of the optimum. Reactions that had zero flux in the FBA analysis (at both optimum values) were considered as potentially invariant between the studied perturbations and thus omitted from the ENZSUB-1 model. The resulting model had 349 reactions involving 267 metabolites. The here-used FBA-based approach for model reduction does not necessarily imply that the eliminated reactions are inactive and that the metabolites involved in these reactions not present in the cell. However, it is assumed that as these reactions are likely to carry very low fluxes under the studied conditions, the associated metabolite pools are likely to be weakly affected due to changes in the fluxes through these reactions. Although this approach is useful, the assumption is not fool-proof as we indeed found certain measured metabolites that were intermediates in pathways with zero fluxes (Pimelic Acid, PIMExt, Myristic Acid, C140xt, trans-4-hydroxy-L-proline, Itaconate, Nicotinate, 4-Aminobenzoate, THMxt). The first six of these metabolites were detected as 'invariant' by the FBA approach due to the fact that that these metabolites are not connected to the overall network (Forster et al., 2003a). However, here we restored back reactions involving these measured metabolites, and the resulting model comprised a total of 285 metabolites participating in 361 reactions (Supplementary Figure S-5.1). Even though certain reactions may be removed from the analysis by using this approach, the algorithm will still correctly identify reporter reactions, given the metabolome dataset. The resulting metabolic net-

work, ENZSUB-2 model, was substantially enriched in terms of the content of measured metabolites (now accounting for about 30%).

In order to further focus the analysis only on reactions involving measured metabolites, ENZSUB-3 was constructed by keeping only reactions that involved at least one measured metabolite. Additionally, only one member of the NADH/NAD^+, NADPH/$NADP^+$, $FADH_2$/FAD^+ cofactor pairs, when available, was retained in the remaining reactions since the levels of members of each pair were assumed to be interdependent. The resulting metabolic network, ENZSUB-3, included a total of 178 metabolites participating in 139 reactions, which corresponds to more than 47% of the available quantitative metabolome data (Supplementary Figure S-5.1). The 139 reactions included in the model are given in the Supplementary Table 1.

The significance of change in the levels of metabolites between any two conditions was calculated by applying a statistical test (see methods section). However, it is difficult to deduce which reactions in the cell are affected most by only judging the significance of the change in metabolite levels, since the number of the metabolic reactions in the cell is high and one metabolite usually appears in more than one reaction. Thus, we calculated a normalized Z-score for each reaction based on the z-values of its neighboring metabolites (p-values of individual metabolites were converted to Z-scores by using inverse normal cumulative distribution function, see methods section). Here we assume that the calculated reaction Z-scores can be regarded as an indicator of the significance of how the reactions respond to the studied perturbation at metabolic level. This assumption is based on the fact that metabolite levels are governed by changes in fluxes and enzyme activities (Nielsen, 2003). Reactions exhibiting significant changes (typically $z > 1.28$, corresponding to $p < 0.10$) for the perturbations analyzed were identified by using the graph representation of the derived metabolic model, ENZSUB-3, and listed in Table 5.1 and Table 5.2. A loose cut-off was deliberately chosen since we did not want to be too-biased in the light of the fact that measurements were not available for all of the metabolites in the model, and thus the resultant p-values are in fact, in general, shifted to high values due to randomly selected p-values for those unmeasured metabolites.

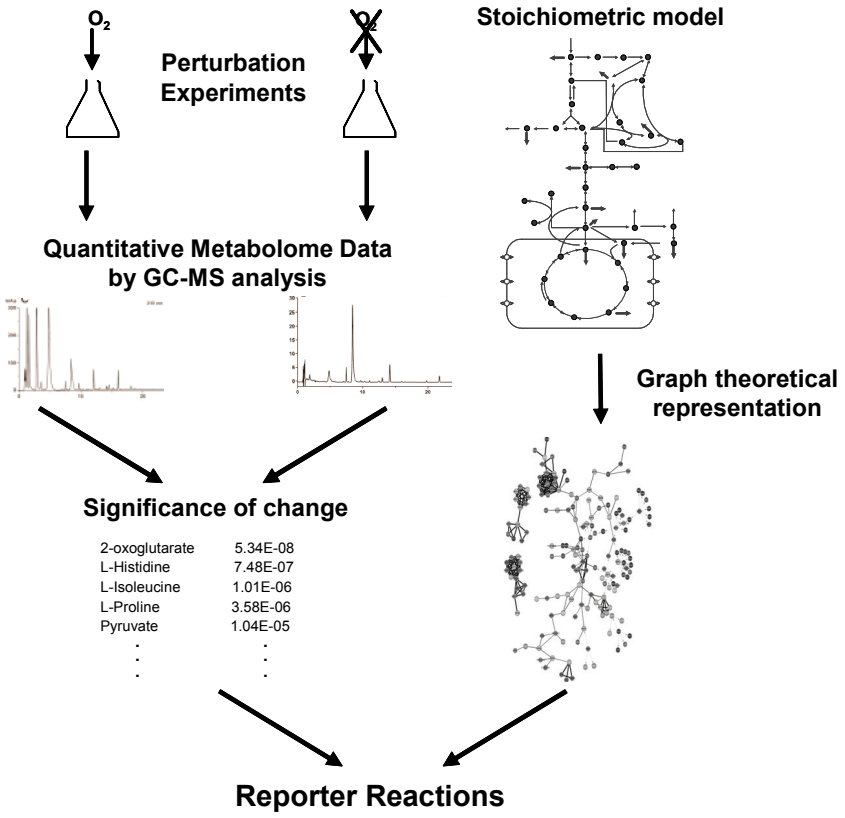


Figure 5.1. Reporter reaction algorithm to identify differential reaction significance by integrating metabolome data with metabolic networks. Quantitative metabolome data obtained from perturbation experiments is interpreted in terms of significance of change, and mapped onto the stoichiometric network which is represented as bi-partite undirected graph, to identify reporter reactions.

Table 5.1. Reactions with significant z-scores ($p < 0.10$, $z > 1.28$) in response to genetic perturbations by altered redox metabolism and environmental perturbation by oxygen availability[a,b,c]. The number of measured metabolites and the total number of metabolites for each reaction are also given in parentheses. The explicit form of the reactions can be followed from Supplementary Table 1.

Genetic Perturbation (aerobic)			**Genetic Perturbation (anaerobic)**			**Environmental Perturbation (wild type strain)**		
VALsyn	(4/4)	2.90	AGX1	(4/4)	2.67	UGA^{ES}	(5/5)	2.41
ALT	(4/4)	2.83	ALT	(4/4)	2.35	ALT	(4/4)	2.34
LEUsyn^{ES}	(5/6)	2.66	PROsc	(2/3)	2.08	AGX1	(4/4)	2.34
TYRsyn	(3/4)	2.54	LEUsyn^{ES}	(5/6)	1.80	CAR2	(3/4)	1.95
CAR2	(3/4)	2.50	ASP3-1	(2/3)	1.78	LEUsyn^{ES}	(5/6)	1.95
PHEsyn^{ES}	(3/5)	2.25	U46_	(3/4)	1.64	TYRsyn	(3/4)	1.92
AGX1	(4/4)	2.01	CHA1p	(2/3)	1.58	VALsyn	(4/4)	1.87
AAT	(4/4)	1.86	PHEsyn^{ES}	(3/5)	1.57	PHEsyn^{ES}	(3/5)	1.74
ILEsyn^{ES}	(6/7)	1.77	PUT1	(2/3)	1.55	SERsyn^{ES}	(4/6)	1.67
SUCsc	(2/3)	1.66	VALsyn	(4/4)	1.54	GAD1	(2/3)	1.47
SDH	(2/3)	1.63	GLY1	(2/3)	1.50	GDH13	(3/4))	1.44
HISsyn^{ES}	(4/10)	1.58	SERsyn^{ES}	(4/6)	1.41	ASP3-1	(2/3)	1.39
ASP3-1	(2/3)	1.57				GDH2	(3/4)	1.38
GDH2	(3/4)	1.55				MYRsc	(2/2)	1.36
DLD	(2/4)	1.51				ILEsyn^{ES}	(6/7)	1.36
UGA^{ES}	(5/5)	1.48				HISsyn^{ES}	(4/10)	1.34
SERsyn^{ES}	(4/6)	1.46				GLYsyn	(2/4)	1.30
LEU4	(2/4)	1.36				U155_	(4/4)	1.29
FUM	(2/2)	1.28						

[a]Reactions specific to each perturbation are given in bold letters. [b]ES means that the corresponding reaction is an enzyme subset consisting of combination of more than one reaction. [c]sc in some of the reaction names stands for 'secretion', indicating that they are secretion reactions.

Table 5.2. Effect of media change (standard medium vs. VHG medium) on each strain analyzed by the developed approach. Reactions with significant z-scores ($p < 0.10$, $z > 1.28$) are shown[a]. z-scores of the gene expression changes are also given for comparison. z_{RE}: z-scores of reactions calculated by the developed approach, z_{GE}: z-scores of genes/gene groups calculated from asscociated p-values from transcriptome data. The number of measured metabolites and the total number of metabolites for each reaction are also given in parentheses.

Media Change for laboratory strain (CEN.PK113-7D)				**Media Change for industrial strain (Red Star)**			
		z_{RE}	z_{GE}			z_{RE}	z_{GE}
ALT	(4/4)	2.50	2.48	ALT	(4/4)	2.48	1.80
AGX1	(4/4)	2.45	0.86	AGX1	(4/4)	2.31	2.21
UGA [ES]	(5/5)	2.18	1.69	UGA [ES]	(5/5)	2.23	0.38
ECM40	(3/4)	1.85	2.39	U155_	(4/4)	2.01	-
GLUsc	(2/3)	1.85	1.17	ASN	(4/7)	1.85	0.54
ASN	(4/7)	1.84	2.30	TYRsyn	(3/4)	1.84	3.41
CAR2	(3/4)	1.74	0.57	GLUsc	(2/3)	1.81	0.79
LYSsyn [ES]	(7/8)	1.67	2.46	PHEsyn [ES]	(3/5)	1.65	0.98
TRP23	(3/5)	1.67	1.31	TRP23	(3/5)	1.65	0.78
ASP3-1	(2/3)	1.47	1.45	PROsc	(2/3)	1.45	0.90
CHA1p	(2/3)	1.47	0.93	ALAsc	(2/3)	1.45	0.75
U42_ -43_	(2/3)	1.47	-	GLYsc	(2/3)	1.45	0.80
ASPsc	(2/3)	1.43	2.09	LACsc	(2/3)	1.45	0.86
PROsc	(2/3)	1.43	1.40	PYRsc	(2/3)	1.45	0.86
ALAsc	(2/3)	1.43	1.90	SUCsc	(2/3)	1.45	-
GLYsc	(2/3)	1.43	1.66	CITsc	(2/3)	1.45	-
LACsc	(2/3)	1.43	0.81	AKGsc	(2/3)	1.45	-
PYRsc	(2/3)	1.43	0.81	U88_	(2/3)	1.43	-
SUCsc	(2/3)	1.43	-	GAD1	(2/3)	1.43	1.21

GLY1	(2/3)	1.41	0.41	ILEsyn^{ES}	(6/7)	1.42	1.90
VALsc	(2/3)	1.38	1.55	ASP3-1	(2/3)	1.41	1.34
PHEsyn^{ES}	(3/5)	1.36	1.66	U42_-43_	(2/3)	1.41	-
GAD1	(2/3)	1.29	1.46	LEUsyn^{ES}	(5/6)	1.38	0.91
				ASPsc	(2/3)	1.29	0.86
				ARG5,6-8ES	(4/8)	1.28	0.88

[a] Reactions specific to each perturbation are given in bold letters.

5.4.2 Effect of an altered redox metabolism and oxygen availability

As a first demonstration of our approach we considered data from metabolome analysis of two different *S. cerevisiae* strains, a wild type laboratory strain (CEN.PK.113-7D) and a redox engineered strain, which was carried out in batch cultures under two different environmental conditions (aerobic and anaerobic) in standard mineral media with glucose as the sole carbon source (Villas-Boas et al., 2005c). The redox engineered strain carrying a deletion of the NADPH-dependent glutamate dehydrogenase encoded by *GDH1* and an over-expression of the NADH-dependent glutamate dehydrogenase encoded by *GDH2* was constructed by dos Santos et al. (2003). Three different perturbations were analyzed here: genetic change under both aerobic and anaerobic conditions (wild type versus redox engineered strain), and environmental change for the wild type strain (aerobic versus anaerobic). Since it was reported that sample-to-sample variability exceeds flask-to-flask variability, replicate samples from different shake flasks were treated equivalently (Villas-Boas et al., 2005c). Accordingly, the metabolome dataset includes around 15 intracellular and 9 extracellular replicates for each experimental condition. The dataset used in this study is available in the supplemental material as normalized abundances of GC-MS peaks.

Comparison of the wild type and mutant strains revealed that the genetic changes do not alter the basic growth characteristics in aerobic (dos Santos et al., 2003) and anaerobic (Nissen et al., 2000) batch cultivations. Our approach, however, captures the associated changes in different cellular pathways by identifying a number of significantly affected reactions due to these perturbations. The detected reactions (Table 5.1) belong to many different amino acid pathways, indicating a widespread effect of the mutation on the cellular metabolism. The present integrated approach also differentiates between the genetic perturbation under aerobic and anaerobic conditions as there are reactions that are specific to each condition.

Genetic perturbations (wild type versus redox engineered) used in the present study are directly related to a changed redox metabolism. Environmental perturbation (aerobic versus anaerobic) is, however, also associated with a changed redox metabolism due to the direct effect of oxygen availability on the operation of the TCA cycle and the pentose phosphate pathway, and hence on the redox state of the cell. This is also reflected in the identified reporter reactions since a number of common significantly changed reactions are observed for the two different types of perturbation (Table 5.1, Supplementary Table 5.2).

The glutamate decarboxylase reaction (GAD1) appears as a significantly changed reaction specific to the environmental perturbation of the wild type cells, which implies a major role of this reaction during respiratory growth (Table 5.1). Indeed, it was reported (McCammon et al., 2003) that the defects in any of the 15 TCA cycle genes, associated with the slowing down of the respiratory metabolism, result in a substantial decrease in the mRNA levels of *GAD1*, which is in agreement with our findings. GAD1 constitutes the first step of the glutamate catabolic pathway towards succinate (Coleman et al., 2001). The downstream steps of the pathway are catalyzed by Uga1p and Uga2p (UGA^{ES}), which are affected most by the environmental perturbation (Table 5.1). Detection of all reactions of this pathway (GAD1, UGA^{ES}) as responsive to the oxygen availability (Figure 5.2a) indicates that they have a key role in succinate production via glutamate under anaerobic conditions where the yeast is secreting succinate. In fact, this pathway was found to be activated during oxidative (Coleman et al., 2001) or osmotic (sugar) (Erasmus et al., 2003) stress to control the redox balance of the cell.

Although the glyoxylate cycle is generally believed to be repressed during growth on glucose, Villas-Bôas et al. (Villas-Boas et al., 2005a) found that an alternative pathway for glyoxylate biosynthesis is active in *S. cerevisiae*. Examination of the Z-scores of reactions involving glyoxylate for all the analyzed perturbations revealed that AGX1 (reaction of enzyme encoded by *YFL030w*), which enables synthesis of glyoxylate from glycine, has much higher scores for all the perturbations compared to the reactions of the glyoxylate pathway (ICL and MLS) (Figure 5.2b). Thus, our analysis supports the presence of an alternative pathway catalyzed by AGX1 leading to the biosynthesis of glyoxylate from glycine.

Reporter reaction analysis also identifies that the genetic perturbation results in metabolic changes around the genes that are perturbed (Figure 5.2c). Thus, the reaction responsible for the over-expressed gene in the redox-engineered strain, GDH2, has a significant Z-score for the genetic perturbation under aerobic condition. It should be mentioned that a genetic perturbation of a gene should not necessarily result in that the corresponding reaction comes out as a reporter

reaction, as certain genetic perturbations may lead to only small changes in metabolite levels. However, in this case there are two genetic modifications around α-ketoglutarate and glutamate (deletion of *GDH1* and over-expression of *GDH2*) which leads to identification of GDH2 as reporter. For the genetic change under anaerobic conditions, the detected significance of GDH2 is comparably lower. However, an indirect effect of the genetic modification in the glutamate biosynthesis can be observed from the presence of transaminase activity associated with some of the identified reporter reactions for this perturbation (conversion of glutamate to α-ketoglutarate by ALT, LEUsyn^{ES}, PHEsyn^{ES}, VALsyn, SERsyn^{ES}, Table 5.1, Supplementary Table 2). On the other hand, the aerobic-anaerobic shift for the wild-type gives rise to nearly the same Z-score for GDH2 reaction as the genetic perturbation under aerobic conditions. One explanation for this similarity in behavior would be that oxygen availability may have a direct effect on glutamate dehydrogenase genes; that is, cessation of oxygen uptake or manipulation of redox metabolism may result in similar effects on this node in the metabolism. In fact, in chemostat cultures, *GDH2* is associated with a significant transcription change when subjected to the same environmental perturbation (Piper et al., 2002). On the other hand, it is not possible to make a definite interpretation about the effect of the mutation on the deleted gene, *GDH1*, by looking at the Z-score of GDH13 reaction since the reaction catalyzed by Gdh1p is identical with that catalyzed by Gdh3p. Consequently, what is reflected by this Z-score is the 'combined' response of these two enzymes. The reason that the GDH13 reaction is not identified as a reporter reaction whereas the GDH2 reaction is identified can only be explained by either a different response in the co-factor level as a consequence of the perturbations, i.e. the NADPH/NADP$^+$ levels do not change as much as the NADH/NAD$^+$ levels, or due to measurement errors of these co-factors (these co-factors are inherently difficult to measure).

Since the TCA cycle activity is known to be low under anaerobic conditions, the associated effect of genetic mutation under this condition is expected to be weaker than the other two perturbations analyzed. The Z-scores for the SDH and FUM reactions (both being part of the TCA cycle) are clearly in agreement with this expectation (Figure 5.2d). These two reactions are also members of the electron transport system, and this further explains why the metabolites surrounding these reactions exhibit remarkably weaker coordinated change in the genetic perturbation under anaerobic condition than in the other perturbations.

Similarly, the Z-scores of key reactions involving oxaloacetate suggest that these reactions are mainly affected in the redox engineered strain under aerobic conditions (Figure 5.2e), and AAT, a transamination reaction leading to the conversion of oxaloacetate to aspartate, appears to be the

key reaction where oxaloacetate is involved. There is no literature data available about the effect of the genetic perturbation on this metabolic reaction but as the genetic perturbation results in a changed ratio of glutamate to 2-oxoglutarate (Villas-Boas et al., 2005c) that may have effected this important transamination reaction.

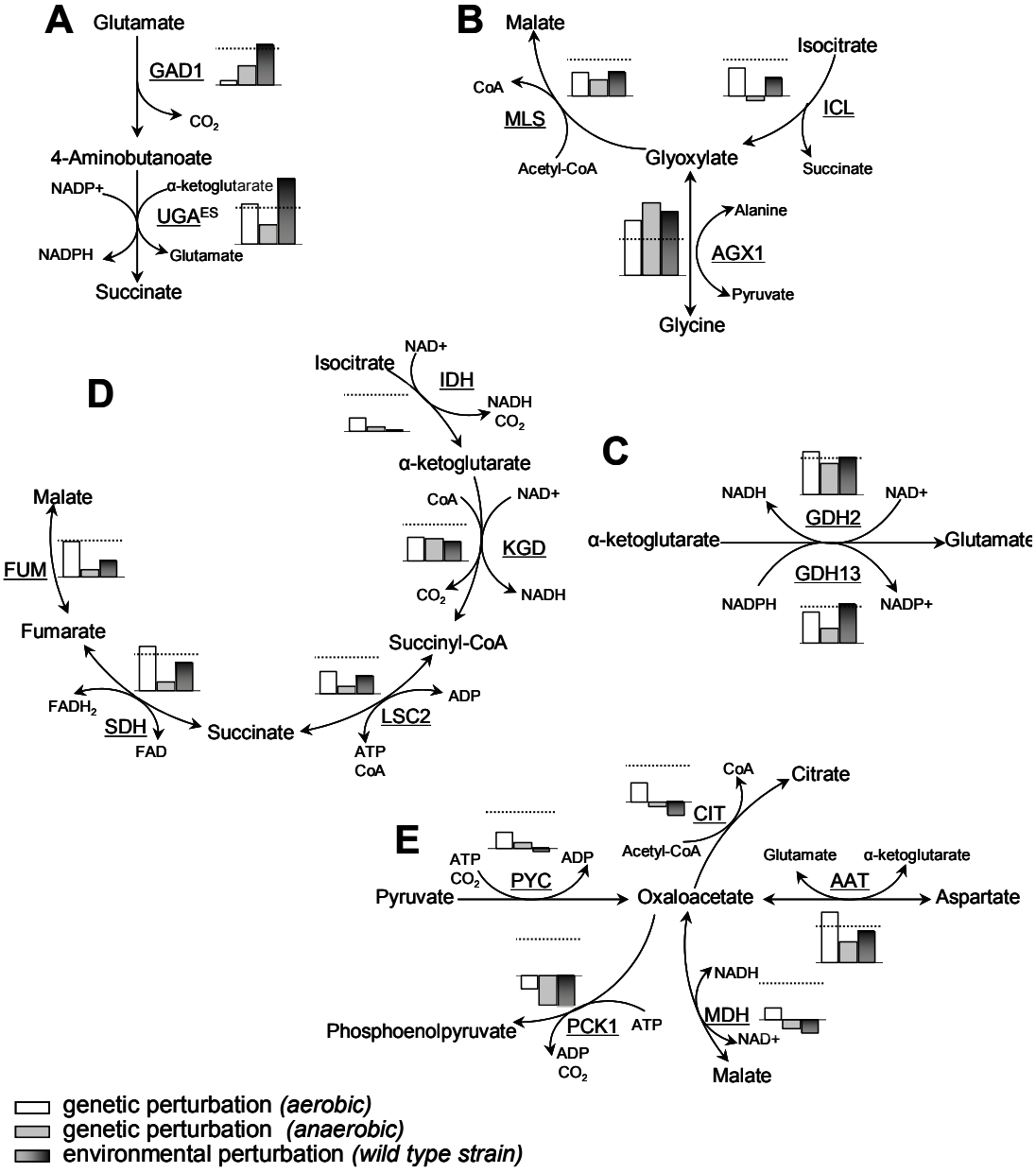


Figure 5.2. Example pathway structures based on Z-scores of reactions, which demonstrate the metabolomic response of the selected reactions in the reported case studies, viz., the effect of an altered redox metabolism and aerobic/anaerobic growth. The dashed lines correspond to the cut-off of 1.28 (p = 0.10). See text for detailed discussion. a) Glutamate catabolic pathway. b) Glyoxylate metabolism. c) Glutamate dehydrogenation reactions. d) TCA cycle. e) Oxaloacetate metabolism.

5.4.3 Effect of very-high-gravity fermentation

As a second demonstration of our approach we used metabolome data from two different *S. cerevisiae* strains, a laboratory strain (CEN.PK.113-7D) and an industrial strain used for fuel ethanol production (hereafter termed as "Red Star"). For both strains the data were obtained from anaerobic batch cultures under two different cultivation conditions; exponential growth in a glucose containing standard mineral media and the stationary phase in a maltodextrin containing very-high-gravity (VHG) mineral media (Devantier et al., 2005a). Environmental perturbations obtained through variation in the media were analyzed here for each strain. The intracellular metabolome dataset includes 4 replicates for the standard medium and 8 replicates for the VHG medium. The extracellular metabolome dataset has 6 replicates for each condition. The complete dataset is available in the supplemental material.

As for the first case study discussed above, the two media perturbations analyzed revealed the same trend for the glyoxylate reactions, pointing to substantial regulation of the AGX1 reaction node in both perturbations (data not shown). In case of the glutamate metabolism, all the reactions have noticeably higher Z-scores, except GDH2, implying that this pathway is highly affected by VHG associated media changes. All of the TCA cycle reactions shown in Figure 5.2d have very low Z-scores, in accordance with the fact that the cycle is barely operational under any of the experimental conditions studied (anaerobic fermentations). For reactions involving oxaloacetate, AAT again appears to play the major role as observed in the first data set, in parallel with the graph shown in Figure 5.2e.

The reaction governed by Gad1p, which catalyzes decarboxylation of glutamate – a reaction that is generally considered to be associated with stress, is found to be significantly changed in both strains when the media was changed (Table 5.2). A noticeably lower score was obtained for comparison of the two strains grown on the standard medium (results not shown), which shows that the standard medium imposes less stress compared with the VHG medium where sugar and ethanol stresses are predominant. The appearance of all reactions (GAD1, UGA^{ES}) involved in the glutamate catabolic pathway as reporter reactions when the media is perturbed (Table 5.2) points to the fact that this perturbation has a major effect on the amino acid metabolism, and probably also on the redox balance in the cell. The results of transcriptome analysis for the same strains in standard and VHG media (Devantier et al., 2005b) indicate that the strains have differences in their redox balancing confirming our finding.

A large number of transport reactions were found to have significant Z-scores (Table 5.2). GC-MS analysis of extracellular metabolites in the VHG medium revealed many more metabolites compared to what is found in the standard medium, explaining the appearance of transport reactions as significant. The here-reported algorithm allowed us to identify and quantify the secretion reactions which are mostly affected from the media change, by integrating both intracellular and extracellular measurements to the reaction network. Secretion of a number of amino acids (glutamate, aspartate, proline, alanine and glycine), and succinate, pyruvate and lactate are commonly and significantly regulated in response to media perturbation for both the laboratory and the red star strain. On the other hand, detection of strain-specific secretion patterns (valine, citrate and alpha-ketoglutarate, Table 5.2) points to differences in operation of the metabolic network in the two strains, possibly arising from the difference in the redox metabolism of the two strains.

Since the change in the fermentation medium led to ethanol and osmotic stress for both strains (Devantier et al., 2005a), it is not surprising that many of the reactions are shared in the identified lists for the two strains in the media comparison (Table 5.2). Transcriptome analysis of this dataset revealed that a substantial part of the significantly changed genes were involved in protein synthesis and amino acid metabolism (Devantier et al., 2005b). Thus, amino acid pathway reactions detected by our analysis (Table 5.2) are in accordance with the transcriptome data. Absence of amino acid synthesis in VHG media due to the cessation of growth in the stationary phase can be a possible cause of the observed differences.

5.4.4 Integration of metabolome data with transcriptome data for understanding regulation

For the latter case study, where the effect of a VHG medium was analyzed on the metabolome of laboratory and industrial strains, there was also performed genome wide expression analysis (Devantier et al., 2005b). This basically enables further evaluation of mode of regulation for the different reactions in the reduced metabolic network. ter Kuile & Westerhoff (ter Kuile and Westerhoff, 2001) introduced the concept of metabolic regulation and hierarchical regulation, where the first indicates that regulation of flux is at the level of enzyme kinetics, i.e. through changes of the metabolite levels, and the second indicates that regulation of flux is at the level of enzyme production/activity (transcription/translation/post-translational modification). As both metabolite data and transcription data are available for this case study, we looked into whether it was possible to identify the type of regulation at the individual reaction level. A major obstacle for this kind of analysis is, however, that we do not have information about changes in fluxes for the analyzed conditions, and such data would also be difficult to obtain. Although there are efficient methods for obtaining data on the metabolic fluxes in the central carbon metabolism

(Nielsen, 2003), it is difficult to get good estimates for the fluxes in all pathways of the metabolic network analyzed here, and even though the fluxes can be calculated by using flux balance analysis, this method is not well suited to give precise estimates for the actual fluxes in networks where there are redundant pathways. In order to proceed with analysis we therefore assumed that whenever there was a coordinated significant change in metabolite levels around a reaction, then it is very likely that the flux through this reaction is also changing. However, there is no guarantee that the flux through this reaction is also changed as there could also be a change in the enzyme concentration, or there could even be altered allosteric regulation of the enzyme, thus keeping the flux unchanged. Thus, our assumption may result in identification of some false positives, but still the analysis would clearly lead to identification of reactions around which there is at least one level of regulation (and possibly several levels of regulation), and we will therefore refer to these reactions as being metabolically regulated. For all the reactions that are not identified as reporter reactions we can not infer anything about whether the flux has changed, but still we can deduce from the transcription data whether there has occurred regulation at the hierarchical level, and even though this does not necessarily mean hierarchical regulation of the flux we will refer to these reactions as being hierarchically regulated. This deduction can still be informative as indicator of the logic of transcriptional regulatory machinery governing gene expression. For cases where there was a significant change at the transcriptional level for an identified reporter reaction we considered this to be a situation where there was mixed regulation.

The metabolic network includes several enzymes (hence reactions) governed by multiple genes. Thus, in order to infer about the significance of change in expression levels for the reactions we summed the transcript levels for all genes coding for the same reaction before applying the statistical test. The p-values of transcripts were then calculated by using a t-test with unequal variance, and further converted into Z-scores to enable a comparison with the Z-scores of reactions based on metabolome data.

Using this approach we grouped all the reactions of the metabolic network into whether they were metabolically or hierarchically regulated (or a combination or not regulated at all) for the VHG dataset. To score the magnitude of the regulation at the hierarchical and metabolic levels we used the corresponding Z-scores. Hereby the qualitative evaluation of Z-scores emerging from the transcriptome and the metabolome data enabled us to get an indication of regulation within the metabolic network (see Supplementary Figure S-5.2, Supplementary Table 3). The cases where only the transcript Z-score is significantly changed can be scored as points with possible hierarchical regulation, whereas the opposite case where only the metabolite based Z-score

has significantly changed implies metabolic regulation of the corresponding reaction (Rossell et al., 2005). When both Z-scores are significant there is regulation shared at both levels, and when none of the Z-scores are significant, it is not possible to infer about at which level there is regulation.

Of the 121 reactions in the model having corresponding genes associated with them, the number of reactions predicted to be regulated hierarchically, metabolically, and at both levels were 56, 7, and 14 respectively for the media perturbation with the laboratory strain, and 31, 14, and 5 for the same perturbation with the industrial strain (Figure 5.3, Supplementary Table 3). For the laboratory strain, 44 reactions were found to be relatively irresponsive to the perturbation. On the other hand, the number of potentially unregulated reactions was much higher (71) for the industrial strain. One explanation for the observed predominance of transcriptional regulation could be the fact that the strains protect themselves against the applied perturbation by mainly changing their gene expression to minimize the changes in the metabolome; an observation also encountered in plants (Hirai et al., 2004). Figure 5.3 and Supplementary Table 3 suggest that metabolic regulation is mainly predominant for secretion reactions and amino acid pathways with or without simultaneous hierarchical regulation, the sole exceptions being proline and methionine/cysteine pathways. It is logical to identify the latter as subjected to different regulation since they are involved in pathways with sulfur assimilation and there were no direct perturbation on sulfur utilization in the experimental study. The type of regulation for a number of reactions differs between the two strains, which supports the finding that gene expression pattern can vary within different *S. cerevisiae* strains (Ferea et al., 1999;Brem et al., 2002;Townsend et al., 2003;Jansen et al., 2005). Ferea *et al.* (Ferea et al., 1999) have reported altered expression levels of genes involved in metabolite transport for strains obtained by adaptive evolution in glucose limited cultures. This observation presents an interesting analogy to our analysis, as the industrial strain is also likely to be a result of adaptive evolution. Similarly, different wild type strains were found to have widespread variations in expression of genes involved in amino acid metabolism (Townsend et al., 2003). In order to further validate that the metabolism is different in the industrial and laboratory strains, we performed principal component analysis of the metabolome data for the VHG medium dataset (Supplementary Figure 3). This shows a clear distinction of the strains indicating that the strains behave remarkably different at the level of metabolome. Our analysis systematically combines the transcriptome and metabolome and deduces the underlying regulation causing these differences in metabolism. Notably, following a change to a high-gravity fermentation medium, transcriptional regulation of metabolism is much more pre-dominant in the laboratory strain as compared to the industrial strain; whereas the number of reporter reac-

tions between two strains is around the same with a 70% overlap (Table 5.2). This strongly suggests that although the industrial strain has a better adaptation of its transcriptional program for high-gravity media, there is still similar metabolic regulation pattern to the laboratory strain. The difference in strains in terms of their response to the same perturbation is, again, very visible in the secretion reactions where laboratory strain attempts to regulate them also at transcriptional level, whereas industrial strain relies predominantly on metabolic control (Figure 5.3, Supplementary Table 3). The lesser degree of transcriptional regulation in the industrial strain could benefit the cells by reducing the investment of resources in transcriptional regulatory machinery.

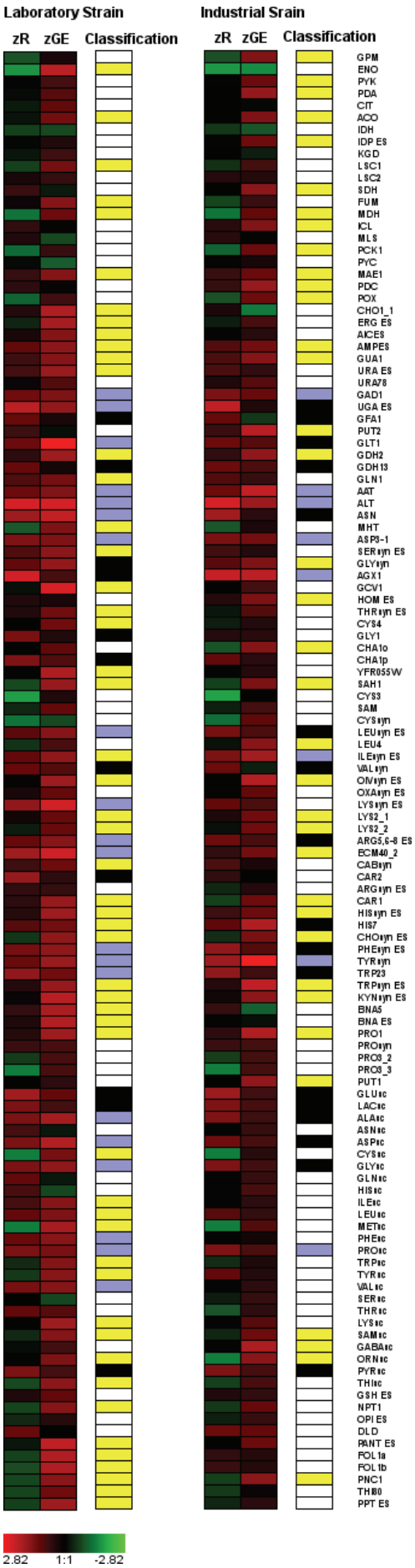


Figure 5.3. Magnitude of the regulation for the reactions of the metabolic network, ENZSUB3, at the hierarchical and metabolic levels for the effect of very high-gravity (VHG) fermentation media on laboratory (CEN.PK113-7D) and industrial (RS) strains. Z-scores calculated based on gene expression changes (zGE) and based on changes in the surrounding metabolites (zRE) are shown. Red means a positive Z-score, and green means a negative Z-score indicating that the regulation is insignificant. Reactions were color-coded with respect to their Z-scores using z = 1.28 (p = 0.10) as the cut-off value to decide on the corresponding regulation type. *yellow:* hierarchically regulation. *black:* metabolically regulation. *violet:* mixed regulation. *white:* statistically insignificant score for both type.

5.5 Conclusions

In the present study, an integrative algorithm based on metabolome data was introduced for the identification of reporter reactions, defined as the reactions that are responding to a genetic or environmental perturbation through a coordinated variation in the levels of surrounding metabolites. We demonstrate that the algorithm functions, even with a small number of measured metabolites (84), which is a typical situation for several currently used technologies. Moreover, the method developed is suitable for mapping the entire alterations associated with a specific perturbation, depending on the advances in analytical detection techniques enabling the measurement of a larger number of metabolites.

Furthermore, when integrated with transcriptome data our approach can be used to infer information about whether a reaction is metabolically regulated or whether it is hierarchically regulated. Our analysis can therefore be regarded as a genome-scale approach towards the integration of different types of omics data by using metabolic networks as a scaffold in order to understand the architecture of metabolic regulatory circuits. Furthermore, our model driven analysis is flexible and will further allow integration of other types of omics data, such as proteomics, and this will further refine the method presented herein to account for the genome-scale alterations in response to genetic as well as environmental perturbations, and hence allow genome-scale identification of regulation in the metabolism.

5.6 Methods

5.6.1 Graph Representation

In the present study, the metabolic network ENZSUB-3 was represented as a bipartite undirected graph to identify reporter reactions. Reactions and metabolites were both taken as nodes, and the edges denoted the interactions between them (Patil and Nielsen, 2005) (Chapter 3). Hence, the graph consisted of 317 nodes.

Different genetic and environmental perturbations associated with the two datasets (Devantier et al., 2005a;Villas-Boas et al., 2005c) were analyzed. The graph representation was used to identify 'reporter reactions' for these perturbations. The algorithm used in the simulations is a modification of the algorithm recently developed by Patil & Nielsen (Patil and Nielsen, 2005) (Chapter 3), which was based on the analysis of transcriptoma data to identify so-called reporter metabolites, the spots in the metabolism with substantial transcriptional regulation. The modified algorithm herein has the capability of identifying reporter reactions, the putative key points in the metabolism in terms of metabolic regulation (Figure 5.1).

5.6.2 Significance Test

The significance of change for the experimental metabolite levels between any two conditions were determined by comparing the levels with the aid of a statistical-test, thereby quantifying the effect of the associated perturbation. For each of the perturbations, the statistical test was applied to the experimental data following the normalization process described by Villas-Bôas *et al.* (Villas-Boas et al., 2005c). Briefly, the normalization process is such that the within-group variances among replicates are reduced and between-group variances are maximized. The Mann-Whitney rank-sum u-test is a nonparametric statistical test which has no *a priori* assumption about the distribution type of the data. It was preferred over the standard t-test since the distribution of levels of some of the metabolites among the replicates, especially NAD+ and NADPH, was found to be skewed rather than normal distributed. The Student t-test assumes normal distribution of the data and compares the mean values whereas the u-test compares medians rather than means. Furthermore, median is a better measure for skewed distributions since it is less sensitive to the extreme scores that can be encountered in the replicates.

5.6.3 Strategy for the Lack of Data

Since the utilized reporter reaction algorithm depends on the scoring of reactions based on the p-values of involved metabolites, the lack of p-values for the 94 metabolites that remain unmeasured in the final ENZSUB-3 model must be handled. Random assignment from GC-MS peaks was used to overcome the problem of the unavailable data. GC-MS spectra contain a large number of unknown peaks due to unmeasured metabolites. All the peaks in GC-MS spectra were deconvoluted for each replicate. The output was normalized by using a Python code which minimizes the sample variability within the classes (Villas-Boas et al., 2005c). Afterwards, the peaks in the spectra within a selected time interval (0.15 minutes) were binned to account for the fluctuations in the retention times using a MATLAB algorithm. This has resulted in the overall detection of 236 unknown peaks for the first dataset (Villas-Boas et al., 2005c), with 116, 178 and 201 non-zero peak comparisons for genetic perturbations under aerobic and anaerobic conditions and environmental perturbations respectively, and 240 unknown peaks for the second dataset (Devantier et al., 2005a) with 129 and 174 non-zero peak comparisons for the environmental perturbation of laboratory and industrial strains respectively. The significance of change for these unknown peaks was quantified for each perturbation by means of p-values using the u-test. These p-values were randomly assigned to the unmeasured metabolites.

5.6.4 Reporter Reaction Analysis

Resultant p-values were converted to Z-scores using an inverse normal cumulative distribution function for further analysis. Each reaction in the constructed graph was scored by calculating the score of the subnetwork formed by its k neighboring metabolites, and z-values of the metabolites were used in the scoring.

$$Z_{reaction} = \frac{1}{\sqrt{k}} \sum Z_{metabolite,k}$$

$Z_{reaction}$ score was then corrected for background distribution using the mean (μ_k) and standard deviation (σ_k) of Z-scores of metabolite groups of the same size, obtained by random sampling from the same metabolic network.

$$Z_{corrected-reaction} = \frac{Z_{reaction} - \mu_k}{\sigma_k}$$

In order to minimize the sensitivity of reporter reactions to the randomly selected p-values for the non-measured metabolites as mentioned above, the reporter-reaction algorithm was executed 1000 times by repeating the random assignment in each case. This repetition eliminated the effect of the p-values of the assigned peaks on results. For each reaction, the Z-scores in each repetition were averaged to get a final Z-score. Those reactions with the highest Z-scores (typically $z > 1.28$, corresponding to $p < 0.10$) can be defined as reporter reactions for a system with complete metabolome data. Since available experimental data were not complete, the calculated Z-scores were used for deducing the relative significance of the reactions in the analyzed perturbations. Namely, we mainly focus on comparative analysis of reactions among the studied perturbations as revealed by Figure 5.2, rather than comparing a reaction to another based on its Z-score. The underlying reason is to avoid potentially incorrect conclusions due to the unmeasured metabolites which have randomly assigned p-values. Additionally, the analyzed reactions have a high percentage of measured metabolite content as indicated in Tables 5.1 and 5.2. In the case of low coverage of measured metabolite content, this method should be followed with caution as the resultant Z-scores of reactions will become insignificant, and such reactions will not be picked up as reporters. However, in future when analytical methods have been further improved it is likely that more metabolites can be measured, and one will overcome this shortcoming and our approach may then be used to infer more solidly about the level of regulation at different parts of large metabolic networks. Based on the features of our algorithm, we suggest certain guidelines for the metabolome measurements in order to effectively exploit our approach: (i) Measurement of me-

tabolites that participate in many reactions (hubs in the metabolic network) will increase the performance of the algorithm, (ii) Measurement of metabolites that participate in certain closely related pathways (metabolites that are closely placed in the network) will increase the confidence in the obtained Z-scores for reactions in those pathways (see supplementary note 1 for further discussion).

5.6.5 Computational Tools

METATOOL 4.3 (Pfeiffer et al., 1999) was used for the identification of enzyme subsets in the UNCOMP model. The codes written in MATLAB 7.0 (MathWorks Inc.) were utilized for the model pre-processing summarized above and to call the algorithm written in C++ for reporter reaction identification. Flux Balance Analysis was performed using in-house software BioOpt employing LINDO API for linear optimization. Deconvolution of peaks in GC-MS spectra for the identification of metabolites based on a metabolite library and for the random peak assignment was achieved using AMDIS software (Stein, 1999), and the peak normalization software was kindly provided by J. F. Moxley.

Acknowledgements: Isabel Rocha (University of Minho, Portugal) is gratefully acknowledged for fruitful suggestions. The authors thank J.F. Moxley (MIT) for the metabolome normalization software. S. Villas-Bôas and R. Devantier are acknowledged for providing detailed information on the experimental data. The authors thank the anonymous reviewers for several constructive suggestions. The research was partly supported by the Boğaziçi University Research Fund through project 04HA502, and by DPT through 03K120250. The doctoral fellowship for Tunahan Çakır is sponsored by BAYG-TÜBİTAK within the framework of the integrated Ph.D. program.

5.7 Supplementary material

Complete supplementary material for this chapter is available online at Molecular Systems Biology journal website (http://www.nature.com/msb). Since most of the data presented in supplementary material is not critical for conclusions drawn, they are omitted here for saving the space. However, two figures that I think will be useful for the understanding of the results are provided below.

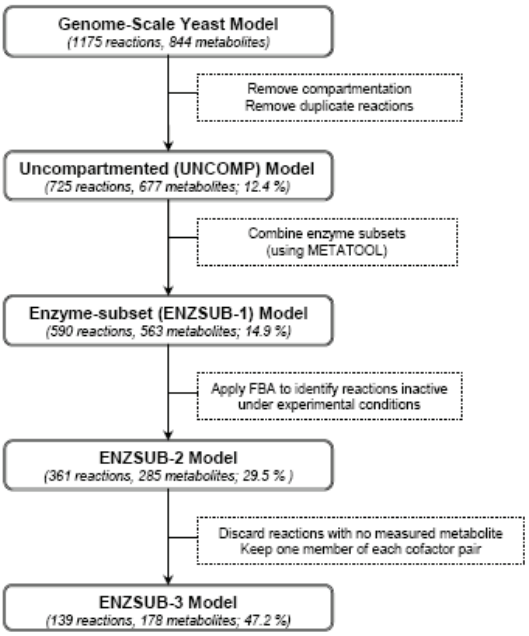


Figure S-5.1. The preprocessing of the model to reduce the fraction of unmeasured metabolites and to focus on reactions involving measured metabolites. Percentages indicate the fraction of measured metabolites in each model.

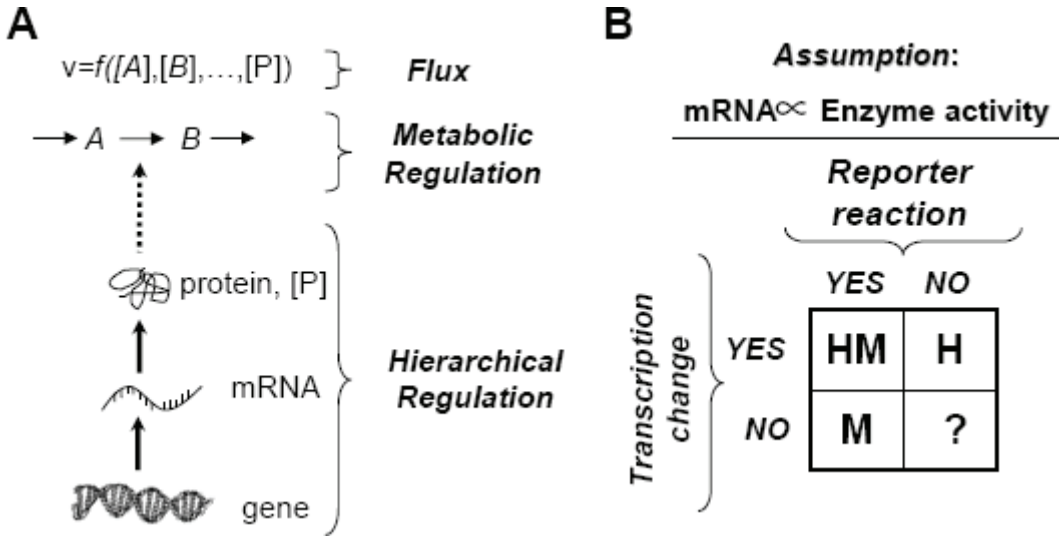


Figure S-5.2. a) Major components of flux regulation. Hierarchical regulation is the regulation imposed by gene-level hierarchy from mRNAs to proteins to enzyme activities. Metabolic regulation is the effect of sub-strate/product concentrations on the fluxes. b) Classification of (reporter) reactions with respect to regulation type. mRNA levels were assumed to reflect enzyme activities.

Chapter 6: Transcriptional Regulation Evolves Around Conserved and Metabolically Related Genes

"Contrariwise," continues Tweedledee, "if it was so, it might be; and if it were so, it would be; but as it isn't, it ain't. That's logic."

6.1 Abstract

Evolution rates of metabolic genes in *Saccharomyces cerevisiae* have been shown to correlate well with their expression levels and functionality (Pal et al., 2006;Pal et al., 2001;Papp et al., 2004;Vitkup et al., 2006;Wall et al., 2005;Dekel and Alon, 2005). However, it is not only the expression of individual genes that dictates the overall functionality of the metabolic network, but also the co-ordinated expression changes in several functionally related genes in response to genetic/environmental stimuli. Regulatory circuits responsible for such orchestrated expression of metabolic genes are not clearly understood in terms of evolutionary and topological principles underlying their emergence. We address this question through systemic analysis of gene modules emerging from the metabolic network topology with respect to their sequence evolution rates, shared promoter sequence motifs and transcriptional co-regulation. We found that the sequence conservancy is significantly over-represented in the gene modules associated with metabolites that are crucial for survival of yeast. We further show that several of these gene modules share sequence motifs in their promoter regions and also show a high degree of transcriptional co-regulation evaluated across a large gene-expression dataset. Our results imply that the topology of metabolic network constraints the evolution of regulatory circuits. In yeast some of these regulatory circuits are built around the evolutionary conserved metabolic neighbors.

6.2 Background

Metabolism plays a central role in the functioning of cells by providing a thermodynamically favourable environment and essential building blocks for growth and maintenance. This pivotal role of metabolism is evident by two facts: i) metabolic pathways are highly conserved across different species ranging from bacteria to humans (Peregrin-Alvarez et al., 2003), and ii) the cellular response to genetic/environmental perturbation is often reflected or mediated through the metabolism (Patil and Nielsen, 2005;Ihmels et al., 2004). Indeed, the evolution of DNA sequences at the level of individual metabolic genes (nucleotide substitution rates amongst the orthologs from closely related species) has been found to correlate well with their functional importance as reflected in the growth phenotypes of the respective deletion mutants (Wall et al., 2005). Moreover, it has also been hypothesized and shown that highly expressed genes evolve more slowly (Pal et al., 2001). These results provide clues as to which genes perform functions that are crucial for the survival and/or adaptation of an organism in light of environmental/genetic challenges faced. Transcriptional regulatory networks aid in adaptation and fine tuning of cellular metabolism in response to perturbations and thus play a key role in evolution and survival of an organism. However, the evolutionary and operational principles underlying the

emergence of these regulatory circuits are still largely unclear. We have previously demonstrated that the operational principles governing metabolic regulation can be uncovered by integrating gene expression data with metabolic network topology (Patil and Nielsen, 2005) (Chapter 3). We found that cells respond to perturbations through transcriptional changes in the metabolism that are centred on certain perturbation specific hot-spot metabolites, termed reporter metabolites. Co-ordinated transcriptional changes around metabolites are indeed necessary for, either to maintain homeostasis or to change the enzyme and metabolite levels so as to adjust to the new flux demands placed on the metabolic network by perturbation (/s). The transcriptional co-regulation of the genes surrounding a metabolite is thus, in part, stoichiometric and thermodynamic necessity. This link between metabolic network topology and transcriptional regulation suggests that the regulatory machinery has evolved for and around the metabolites. Here we test this hypothesis from the metabolite-centred systems perspective and investigate the relationship between sequence evolution rates and the architecture of regulatory circuits (Figure 6.1).

6.3 Results and Discussion

We used the sequence evolution rates for *S. cerevisiae* (Pal et al., 2006;Kellis et al., 2004) metabolic genes in order to identify the metabolites around which the collective evolution rate of corresponding enzyme coding genes (hereafter also referred to as metabolite's neighbour genes) is significantly lower (Conserved-Reporters) and significantly higher (Changed-Reporters) (Methods). Thus Conserved/Changed Reporter metabolites represent evolutionary hot-spots in the metabolic network, with Conserved-Reporter metabolites representing spots in the metabolism around which there has been a high degree of sequence conservation (Figure 6.2). Enzymes surrounding the metabolites involved in the upper part of glycolysis and storage carbohydrate metabolism show significant conservancy together with mitochondrial NADH, H+ and Coenzyme-A; marking their importance in the survival and normal functionality of yeast metabolism. Conservancy around metabolites may be either due to their critical role in survival and/or due to sequence-optimal nature of operation of the surrounding enzymes. Indeed several of these metabolites are located around branch-points of central carbon metabolism where precursors for biomass synthesis are drawn (Figure 6.2). On the contrary, Changed-Reporters mainly reside in the pathways that are either relatively infrequently used (e.g., utilization of alternative carbon sources) or are probably still in a sequence-sub-optimal region of activity, and thus the rapid sequence evolution is directed towards optimizing the activity of the corresponding enzymes. These observations marks the close relationship between the stoichiometry of cellular metabolism, as partially reflected in the metabolic network topology, and the evolutionary changes shaping the function-

ality of this network. Notably, it is possible to uncover these evolutionary hot-spots only through holistic analysis of functionally related gene modules (here defined by using metabolic network topology) and not by considering evolution at the level of individual genes.

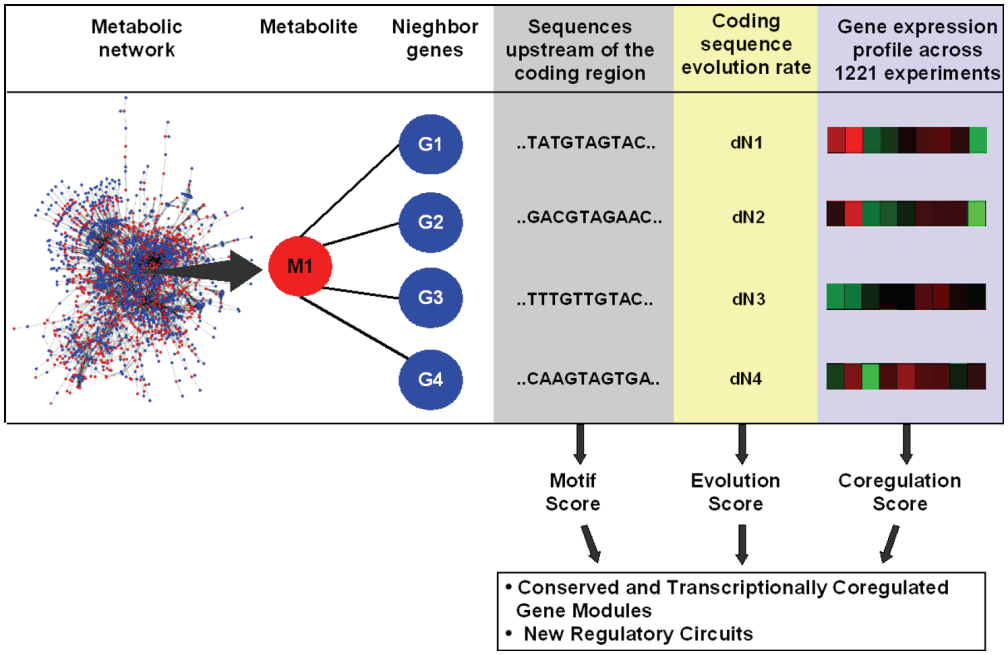


Figure 6.1. Schematic representation of the methodology used to uncover the evolutionary principles underlying the transcriptional regulation observed in yeast metabolism. For each metabolite from genome-scale metabolic model of *S. cerevisiae* we estimated the motif score (signifying the likelihood of observing a common sequence motif in the promoter region of the metabolite's neighbour genes), evolution score (signifying the collective conservancy of the metabolite's neighbour genes) and co-regulation score (signifying the extent of transcriptional co-regulation among the metabolite's neighbour genes). These three parameters lead to the identification of gene modules that are functionally related via metabolic network topology and show significant transcriptional co-regulation across a large gene expression dataset. Identified motifs in the promoter regions of these genes represent known/potential binding sites for transcriptional factors regulating their expression.

Metabolite's neighbour genes can be viewed as functionally related gene modules that naturally emerge through metabolic network. These modules allowed us to extend the gene expression analysis from individual gene level to the transcriptional co-regulation across all of the module genes. To this end, we compiled a transcriptome dataset for *S. cerevisiae* consisting of 1221 public domain experiments. This dataset spans transcriptional response of yeast metabolism to a variety of genetic (~944) and environmental (~277) perturbations. We calculated a score for each metabolite attesting the significance of correlation amongst the corresponding neighbour genes (Methods). This score indicates the extent to which the neighbour enzymes of a metabolite are

transcriptionally coregulated as opposed to the correlation amongst the other genes. Over 80 metabolites were found to have highly significant co-regulation (p-value cut-off corresponding to False Discovery Rate of 1) around them (Supplementary Information), supporting our hypothesis of metabolite-centred transcriptional regulation. These metabolites span several different metabolic pathways and include highly connected nodes (hubs) such as NADPH, AMP and ATP, indicating the connectivity independent nature of transcriptional co-regulation centred on metabolites.

Next, we tested whether the remarkable degree of global transcriptional co-regulation observed around certain metabolites could partially be explained through possible regulatory circuits that may be operating via binding in promoter regions of corresponding neighbour genes. We searched for the presence of conserved motifs in the upstream regions (800 base-pairs upstream of the start codon) of a metabolite's neighbour genes and assigned a significance score (motif-score) to each metabolite based on the E-value of the discovered motif (a high motif-score signifies a strongly conserved motif) (Methods). It should be noted that this motif search was performed without using any *a priori* information about the binding motifs of known transcription factors in yeast. This way we could identify putative sequence motifs for promoter binding in the upstream regions of around 87 metabolic gene modules.

Metabolite connectivity and the significance scores of associated gene modules based on sequence conservancy, transcriptional co-regulation and shared promoter motifs present a holistic picture of how individual nodes of metabolic network have evolved to share their functionality and thus bestow operational optimality to the whole network. Indeed, we found that the genes around many of the Conserved-Reporter metabolites are not only highly transcriptionally coregulated but also have significant common motifs in their promoter regions (Figure 6.2). On the other hand, genes around Changed-Reporter metabolites are, in general, weakly coregulated and only few of them show a good motif-score. For certain Conserved-Reporters we could also identify transcription factors that bind to the corresponding high scoring motifs (Methods). This analysis revealed novel regulatory targets for glucose-repression related transcription factors Mig1 and Rgt1 which we hypothesize to be involved in transcriptionally regulating neighbouring enzymes of glucose. This hypothesis was verified by analyzing transcription data from a *MIG1* deletion mutant and a reference yeast strain. Most of the genes identified here to be regulated by Mig1 have significantly altered expression levels even when compared against known regulatory targets of Mig1 (Figure 3). In case of Rgt1, amongst the genes that are known to be regulated by Rgt1, only the neighbour genes of glucose have significantly altered expression in the RGT1 dele-

tion mutant (Table 6.1). Thus, we successfully identified two important elements of the yeast glucose repression regulatory cascade and their regulatory targets solely by using the DNA sequence data and the topology of mass balance network. Furthermore, our hypothesis driven findings surpasses the current knowledge about the regulation by these proteins in terms of correctly predicting the expression changes in their target genes after deletion of the corresponding genes. Although in the case of other putative motifs identified in this study, no known transcription factors could be directly associated with them, they certainly do represent a starting point for unravelling new regulatory circuits in yeast metabolism.

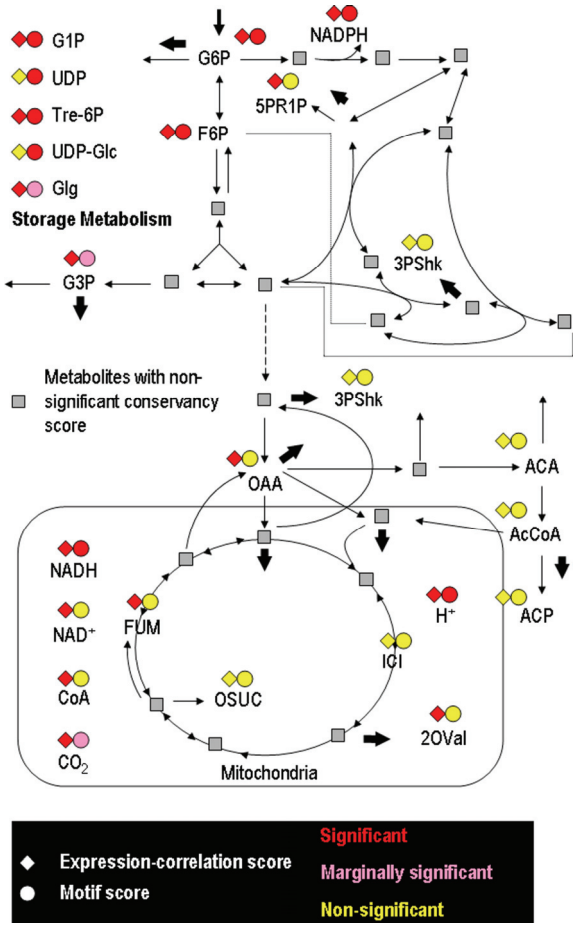


Figure 6.2. Overview of sequence conservancy, motif significance and transcriptional co-regulation in yeast central carbon metabolism. Metabolites with significant conservancy of neighbouring genes are depicted with abbreviated names (Supplementary material). Block arrows indicate the branch-points for biomass precursor molecules. Several metabolites around these branch-points display significant sequence conservancy, thus highlighting there importance for growth and normal cellular functionality. Important examples include metabolites involved in storage carbohydrate metabolism and respiration. Remarkably, metabolites with high conservancy score also show significant shared motifs in the promoter region and strong transcriptional co-regulation on

global-scale. Consequently, we hypothesize that the transcriptional regulation in metabolism has evolved around conserved and functionally related (via network topology) genes.

Table 6.1. Genes known to be transcriptionally regulated by *RGT1*. List of genes is obtained from Yeast Proteome Database (https://www.proteome.com/proteome/YPD). P–values are calculated by comparing expression levels of genes in a reference and *RGT1* deleted strain (Kaniak et al., 2004). Confirming to our hypothesis (see main text), only genes that are functional neighbours of glucose show significant change in expression level.

Gene	Name	p-value	Glucose neighbour?
YHR094C	HXT1	0.0049667	Yes
YMR011W	HXT2	0.0012873	Yes
YHR092C	HXT4	0.014453	Yes
YGL062W	PYC1	0.83389	No
YIL162W	SUC2	0.12378	No

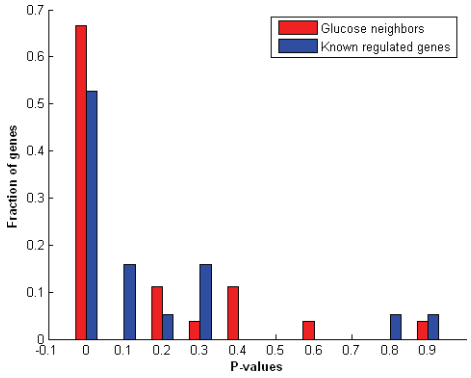


Figure 6.3. Comparative histograms of p-values for genes that are neighbours of glucose and genes that are known to be transcriptionally regulated by Mig1p (YPD- https://www.proteome.com/control/tools/proteome). P-values were calculated by comparing expression levels of genes in a reference and *MIG1* deleted strain (Westergaard et al., 2006). This comparison validates the hypothesis that Mig1p regulates expression of genes encoding the neighbouring enzymes of glucose. The hypothesis was generated through integrative analysis of sequence conservancy, shared promoter sequence motifs and tight regulation of these genes in a large transcriptome dataset covering more than 1200 experiments. Topology of metabolic network around glucose has thus guided the emergence of regulatory circuits for glucose repression.

6.4 Conclusions

Collectively, our results imply that some of the transcriptional regulatory circuits for metabolism are built around the evolutionary conserved metabolite neighbours. This observation indeed rationally explains the emergence of regulatory cascades, as only the conserved and functionally related set of genes would have enough evolutionary time and selection advantage to develop prevailing regulatory circuits around them. Consequently, it appears that the topology of metabolic network not only constrains the flow of material and energy through the cell, but also places functional constraints on the evolution of it's elements and regulatory circuits coordinating their expression levels in response to environmental/genetic stimuli. For *S. cerevisiae*, storage-carbohydrate metabolism, glycolysis and mitochondrial energy generation appears to be under evolutionary selection process for optimization through development of transcriptional regulatory mechanisms for control of their operation. This conjecture is in agreement with a recent study where transcriptional co-regulation of storage-carbohydrates and mitochondrial respiration were found to be prevalent and temporally synchronized (Tu et al., 2005).

6.5 Methods

Conserved/Changed Reporter metabolites

The genome-scale model of *Saccharomyces cerevisiae* (Forster et al., 2003a) was first converted to a bi-partite graph where each metabolite is connected to genes that encode for enzymes catalyzing the reaction involving that metabolite (hereafter also referred to as neighbour genes of a metabolite). Each gene in this graph was then assigned a score as inverse normal cumulative of it's sequence evolution rate (Kellis et al., 2004;Wall et al., 2005), $Z_{evo}^{gene,i}$. A score for each metabolite was then calculated and corrected for background distribution of evolution rates as:

$$Z_{evo}^{metabolite,\mathrm{j}} = \frac{\frac{1}{k}\sum_{k} Z_{evo}^{gene,\mathrm{i}} - \mu^{k}}{\sigma^{k}} \qquad (1)$$

Where μ^{k} and σ^{k} denote the mean and standard deviation of average evolution scores for 10000 randomly selected groups of *k* genes. Scores of metabolites were converted to p-values by using normal cumulative distribution function. Metabolites with high positive scores (low p-values) indicate significant concentration of sequence conservancy in their neighbour genes, while metabolites with high negative scores indicate significantly high rate of sequence evolution around them.

Motif-score for metabolites

Promoter regions (800 basepairs upstream of start codon, downloaded from SCPD-http://rulai.cshl.edu/SCPD/) of neighbour genes of a metabolite were searched for a conserved motif by using RSA-tools (http://rsat.scmbb.ulb.ac.be/rsat/). Corresponding motifs with lowest Expectancy-value (E) were used to assign motif-score to each metabolite as:

$$Z_{motif}^{metabolite,\,j} = CDF^{-1}\left(1 - motif\ Evalue\right) \qquad (2)$$

High motif-score indicates that the found motif is unlikely to be a random event.

Transcriptional co-regulation score for metabolites

In order to quantify how tightly neighbour genes of a metabolite are coregulated under different conditions we compiled a large gene expression dataset consisting of 1221 different experiments available in public domain. For each set of neighbour genes we then calculated all against all Pearson correlation coefficients (P) and converted to Z scores by using inverse normal cumulative distribution function.

$$Z_{pear} = CDF^{-1}\left(P\right) \qquad (3)$$

Significance of scores for each metabolite in contrast to the background distribution was estimated as:

$$Z_{GE}^{metabolite,\,j} = \frac{\frac{1}{n}\sum_{n} Z_{pear} - \mu^{n}}{\sigma^{n}} \qquad (4)$$

Where μ^{n} and σ^{n} denote the mean and standard deviation of average Z scores for 10000 randomly selected groups of *k* genes and n = (k^2-k)/2. Metabolites with high scores imply significant transcriptional correlation amongst neighbour genes.

Transcription analysis of *MIG1* deleted strain

A *MIG1* deleted strain and a corresponding reference strain were both grown in triplicate batch cultures at glucose repressed conditions as described elsewhere (Westergaard et al., 2004). Genome-wide transcription measurement was performed as described elsewhere (Westergaard et al.,

2004). The complete gene expression dataset from this analysis is available in supplementary material.

Scores for all metabolites are available in the supplementary material.

Acknowledgements: We thank J.M. Otero, L. Albertsen, T.L. Petersen and U. Mortensen for comments on manuscript.

6.6 Supplementary material

Complete supplementary material for this chapter is available online at:

http://www2.cmb.dtu.dk/additional_material_for_publications/papers/7/

Since most of the data presented in supplementary material is not critical for conclusions drawn, they are omitted here for saving space.

Chapter 7: Highly connected metabolites harbor significant transcriptional co-regulation

Manuscript describing the results in this chapter is under preparation.

"When I use a word," Humpty Dumpty said, in rather a scornful tone, "it means just what I choose it to mean--neither more nor less." "The question is," said Alice, "whether you can make words mean so many different things." "The question is, "said Humpty Dumpty, "which is to be master--that's all."

7.1 Introduction

Complex cellular operations are extensively reprogrammed at the level of gene expression following environmental/genetic changes. This fine-tuning is often featured by coordinated expression changes in a large number of metabolic genes (Patil and Nielsen, 2005;Ihmels et al., 2004;Daran-Lapujade et al., 2004;Ferea et al., 1999). Transcriptional regulation largely decides which of the large number of possible metabolic states is observed under given conditions, and how this state is altered following a perturbation (/s). Such regulatory responses aid in rapid and global adaptation of the metabolism to altered demands on the synthesis of precursor molecules that are vital for growth and other cellular functions. Thus the transcriptional changes in metabolic networks have been found to be directed towards achieving condition-specific metabolic functionality (Ihmels et al., 2004;Ihmels et al., 2002). Specific metabolic pathways (e.g. as defined in biochemistry textbooks) have been found to exhibit significantly higher transcriptional co-regulation compared to randomly chosen set of metabolic genes (Kuffner et al., 2000;Ihmels et al., 2004;Nacher et al., 2006). This observation highlights the necessity of co-regulation amongst the genes that contribute towards a particular metabolic function. Metabolic pathways can therefore be regarded as hierarchically organized co-regulated modules (Ihmels et al., 2002;Segal et al., 2003). Modular co-regulation of metabolic genes not only explains principles that underlay the complex regulatory machinery orchestrating gene expression but also helps in relating these principles across species (Tirosh et al., 2006).

The pathway based gene modules, however, are derived based on human categorization of enzymes into different pathways. Although pathway oriented classification represent some metabolic functionalities, they do not cover the whole range of metabolic states that a given network can assume (Schuster et al., 2000;Klamt and Stelling, 2002). Moreover, pathway based analysis (and many other network-based studies) often overlook the high connectivity in a metabolic network as several highly connected metabolites are absent from the analysis (e.g. (Vitkup et al., 2006;Kharchenko et al., 2005;Nacher et al., 2006;Ihmels et al., 2004)). Consequently, regulatory principles extracted by these methods have a bias towards local network structures or rigidly defined pathways. . In a complementary approach, we have previously shown that the transcriptional changes in metabolic network are significantly concentrated around experiment-specific metabolites (termed Reporter metabolites) (Patil and Nielsen, 2005) (Chapter 3). Reporter metabolites were found to mark the spots in the metabolism where (known/unknown) perturbation was introduced. Our algorithm does not make any *a priori* assumptions for metabolite connectivity or pathways definitions. Notably, for some of the perturbations, significant expression

changes were centered on highly connected metabolic co-factors (e.g. ATP, NADH). Here we further investigated whether the metabolite centered organization of transcriptional regulation can lead to general principles of metabolic regulation when a large transcriptome dataset (over 1200 experiments for yeast *Saccharomyces cerevisiae*, see Methods) is integrated with network topology.

7.2 Methodology

Figure 7.1 shows a schematic outline of our approach. First we compiled a transcriptome dataset comprising of more than 1200 experiments (Methods). For each of the yeast metabolites we calculated two scores (reporter scores): i) individually for each of the experiments (differential score) based on fold changes in expressions relative to the control, and ii) based on correlations across the whole dataset (multi-dimensional score) (Methods). Reporter scores signify the transcriptional co-regulation observed in neighboring genes of that metabolite as compared to randomly selected genes. Furthermore, we also calculated additional scores by only considering the neighbors that contribute in either production or consumption of that metabolite. In case of multi-dimensional data, further enrichment of biological information was achieved by: i) accounting for the direction of correlation (positive/negative) and ii) additionally considering the correlation across the producing and consuming neighbors (Figure 7.2). Together these scores provide detailed insight into whether the transcriptional changes are directed towards increased (/decreased) consumption (/abundance) of a particular metabolite. In case of differential analysis 87 % of all metabolites were identified as reporter in at least one of the experiments. This result implies that the reporter metabolites are perturbation-specific and several of these perturbations are covered in the compiled dataset. Perturbation specificity of reporter metabolites is further endorsed by the fact that the minimum of average ranks of metabolites over all experiments is very high (about 342) (Figure 7.3).

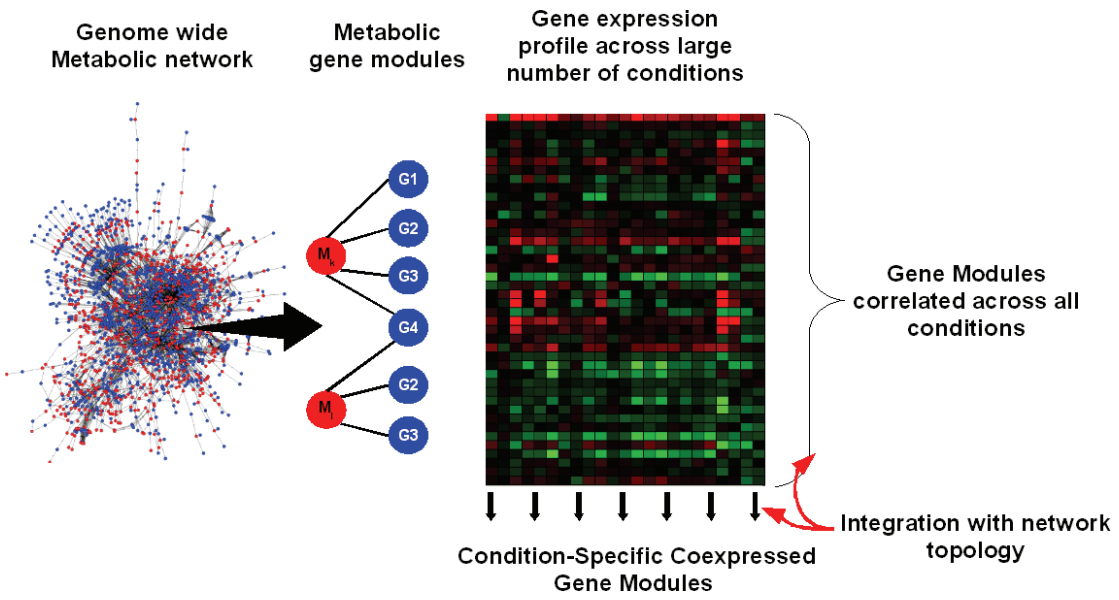


Figure 7.1. Outline of methodology used in this study. Metabolic gene modules emerging from network topology were investigated for significance of transcriptional co-regulation. A large gene expression dataset was compiled from publicly available transcriptome studies. In differential analysis, metabolites marking significant regulation in each experiment were identified. To understand co-regulation patterns across all experiments, a Pearson correlation coefficient based metric was used to identify metabolites for which neighbor genes show strong co-regulation compared to background. Together, this analysis identifies hot-spots in metabolism where significant transcriptional regulation is observed.

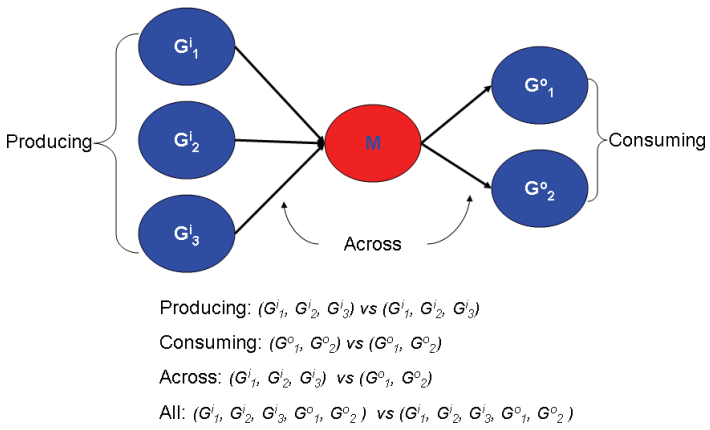


Figure 7.2. Schematic representation of possible transcriptional regulation patterns around a metabolite. Correlation may exist between genes coding for metabolite producing enzymes, metabolite consuming enzymes, across consuming and producing genes or among all of them. First absolute correlation coefficients were used. Further insight into biological meaning of each of these categories was then obtained by considering only positive/negative correlations.

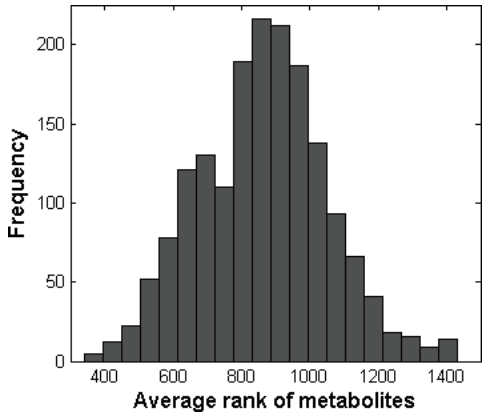


Figure 7.3. Distribution of average rank of metabolites in 1226 experiments analyzed. Each metabolite was assigned three different ranks based on the significance of transcriptional co-regulation observed in the consuming, producing or all of the neighbor genes of that metabolite. Higher the significance, lower the rank.

7.3 Results and discussion

7.3.1 Metabolic regulation: role of pathways and metabolites

Extent of co-regulation observed amongst the neighbor genes of metabolites was found to notably exceed the co-regulation observed in genes involved in common metabolic pathways (as defined in the KEGG database) (Figure 7.4). This contrast is further enhanced when fine-tuned reporter scores with directionality information are considered. Transcriptional co-regulation in metabolic networks thus clearly extends beyond the pathway definitions and more closely intertwined over the network topology. Several metabolites, especially ones with high connectivity, usually span many pathways and act as connecting bridges across these pathways. Consequently, pathways as a whole are not subjected to strict stoichiometric/thermodynamic constraints on their own. Constraints on a pathway can thus only be invoked in the connection with other connected pathways due to overlap of metabolites across pathways. On the contrary, coordinated transcriptional changes around metabolites are indeed necessary for one of two reasons. Either to maintain homeostasis or to change the enzyme and metabolite levels so as to adjust to the new flux demands placed on the metabolic network by perturbation (/s). Thus the transcriptional co-regulation of the genes surrounding a metabolite is, in part, stoichiometric and thermodynamic necessity.

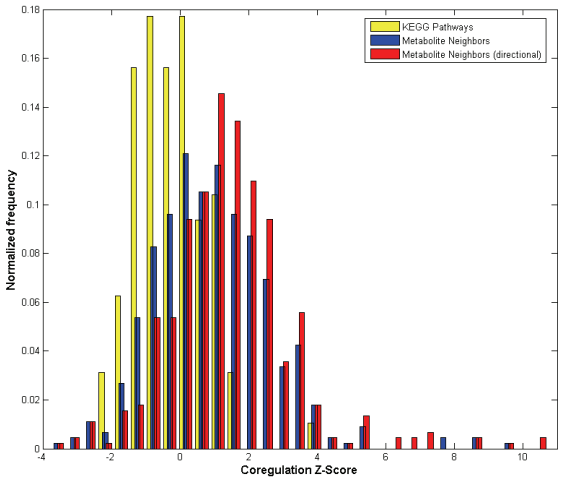


Figure 7.4. Comparison of strength of transcriptional co-regulation in metabolic gene modules based on KEGG pathways and metabolite neighbors. Metabolite neighbors display significantly higher co-regulation Z-score.

7.3.2 Metabolic hubs contribute towards regulation

We further examined whether the degree of correlation around metabolites observed in multi-dimensional analysis (quantified as reporter scores) shows any bias towards high/low connectivity. Remarkably, the distribution of reporter metabolites (154 in total) as compared to non-reporter metabolites (294 in total) shows a shift towards high degree (higher number of neighbor genes) (Figure 7.4). This finding marks the necessity of accounting for the transcriptional correlation linked to highly connected metabolites, which are currently omitted *a priori* from the analysis. Indeed, highly connected metabolites (e.g. NADH, NADPH, ATP and AMP) glue different parts of metabolism together and play a crucial role in determining phenotype by constraining metabolic fluxes based on availability of corresponding metabolite pools. A well known example is formation of glycerol and ethanol in yeast when respiration is limited or compromised and excess NADH is needed to be re-oxidized (Stephanopoulos et al., 1998). Similar phenomenon also has a physiological relevance in humans; such as lactic acid formation in muscles following exercise/oxygen limitation and physiological effects of beta-oxidation of fatty acids leading to rapid NADH consumption. Recent study also indicates that NAD(P)H may play a key role in lipid-induced impairment of glucose-stimulated insulin secretion (Boucher et al., 2004). Predominance of transcriptional control harbored at highly connected metabolites (hubs in metabolic network) can be attributed to stoichiometric/thermodynamic constraints over the whole network. As for individual metabolites, the metabolic network in total is governed by stoichiometric/thermodynamic constraints. Since the "hub" metabolites significantly contribute in keeping the network together (Albert et al., 2000), regulation of consumption/production of these me-

tabolites also plays a significant role on global scale. Although how such regulatory control is mechanistically linked to metabolites is not clear for all such metabolites, there are several examples where metabolic co-factors are directly involved in regulating the expression of several genes, e.g. NAD+ dependent regulation of genes in yeast (Zhang et al., 2002;Lin et al., 2000;Anderson et al., 2003), human (Rutter et al., 2001;Agarwal and Auchus, 2005) and gram-positive bacteria(Brekasis and Paget, 2003).

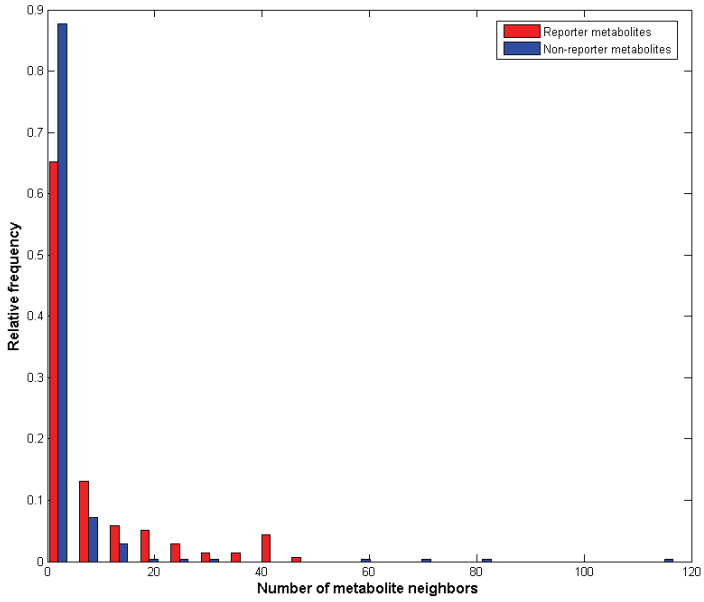


Figure 7.4. Relative frequency of metabolites with different number of neighbors as reporter metabolites. Reporter metabolites are regulatory hot-spots in metabolic networks where transcriptional response is significantly concentrated. Distribution shows a bias of reporter metabolites towards higher number of neighbors.

7.3.3 Distinct response to environmental and genetic stimuli

As in case of multi-dimensional analysis, several highly connected metabolites were also identified as reporters when each of the 1236 experiments were analyzed individually (differential reporter metabolites) (Figure 7.5). Differential reporter metabolites reflect significant fold changes in the expression of the neighboring genes in an experiment and thus provide a unique metabolic signature for each experiment. We classified each of 1236 experiments as either genetic or environmental perturbation. This classification allowed us to investigate whether certain reporter metabolites significantly mark genetic/environmental response of yeast. Metabolites involved in storage carbohydrate metabolism (Glycogen, Trehalose, UDP Glucose etc.) and H2O2 were found to be predominantly marking transcriptional response to environmental perturbations. Storage or utilization of glycogen and trehalose is one of the primary responses of yeast cells to

environmental stress and this phenomenon is clearly reflected at transcriptional level. Moreover, alternate storage and utilization of these carbohydrates has been recently postulated to be a part of natural cyclic operation of yeast metabolism that is closely linked to the cell cycle(Tu et al., 2005). Appearance of H2O2 consumption as another marker for environmental perturbation implies that oxidative stress is often an important feature of environmental stress studied at the transcriptional level. Genetic perturbations, on the other hand, are characterized by significant expression changes in the genes surrounding NADH, NH3, L-Aspartate, Orotidine 5'-phosphate and UMP among others. Genetic perturbations thus seem to be exerting strong transcriptional regulation on redox, nitrogen and pyrimidine metabolism. Identification of NADH as marker for genetic perturbations is particularly interesting and may imply that the cells are attempting to cope with unexpected metabolic perturbations resulting form genetic changes by regulating NADH related genes. This response may achieve two purposes. First is the regulation of metabolism on the global scale, as NADH connects several parts of the metabolism. Secondly regulation of NADH will ensure that the cells are not additionally exposed to deleterious oxidative stress that may result due to unexpected metabolic changes resulting from genetic impairments. Changes in nitrogen and pyrimidine metabolism may be an indirect effect of NADH, as these pathways are closely linked with NAD biosynthesis. The complete list of metabolites marking each of the studied perturbations is available as supplementary material.

7.3.4 Contrasting differential and multi-dimensional analysis

Differential reporter metabolites imply significantly higher changes in the expression of the neighboring genes. On the contrary, reporter metabolites identified from multi-dimensional analysis mean significant correlation between the neighbor genes irrespective of their expression levels. These two classes of metabolites convey different information in terms of expression levels and regulation. To gain insight into these differences we compared metabolite scores in multi-dimensional analysis against number of individual experiments in which those metabolites were classified as reporter (in differential analysis). Metabolites that scored high in multi-dimensional data but were infrequently identified as reporter in differential analysis imply high transcriptional correlation amongst the neighbor genes but relatively low fold changes in the expression levels. Several of the genes associated with the metabolites in this category (FADH2, Orthophosphate, ATP and Fumarate) are mitochondrial and are related to respiration. Their characteristic expression pattern suggests that the functional changes in the enzymes around these metabolites are achieved with low but coordinated changes in expression levels. This may be a result of potential flux limitation at the level of substrate availability.

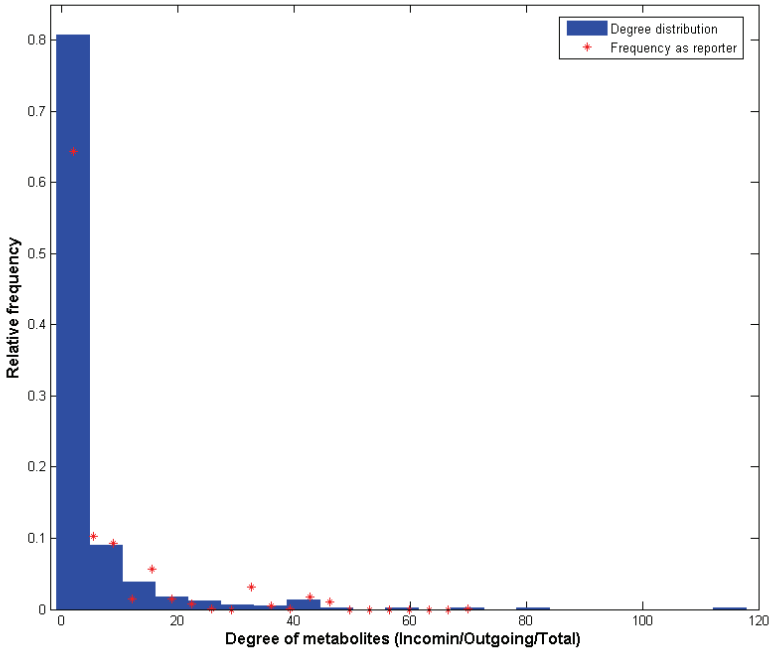


Figure 7.5. Distribution of degree of metabolites in metabolic network of yeast. The histogram is overlaid with frequency distribution of metabolites as reporter.

Another class of metabolites was defined as those that scored low in multi-dimensional analysis but were relatively frequently identified as reporter in differential analysis. Large fold changes but insignificant correlation around these metabolites hints towards redundant functionality and "more the better" relation between expression level and function. Indeed, neighbor genes of metabolites in this class are by and large amino acid transporters which concur with these characteristics.

7.3.5 Isozymes and highly connected metabolites

Over 20% of the metabolic reactions in yeast are catalyzed by more than one enzyme. Significance of these iso-enzymes is generally attributed to vital essentiality of corresponding enzymes for growth during evolution (Papp et al., 2004). We investigated whether this functional complementation through iso-enzymes is reflected in their expression. To this end we calculated the significance of correlation among the iso-enzymes across all experiments under study. Around one-third of 108 iso-enzymes showed significant correlation. Several iso-enzymes thus appear to share similar regulatory mechanisms for their expression. Notably most of the iso-enzyme correlations are positive. Positive correlation among iso-enzymes implies that, either flux through the

corresponding reaction is limited by enzyme availability at transcription level, or that iso-enzymes act as buffer against genetic disturbances such as mutations or deletions. This conjecture is in agreement with the hypothesis proposed by Ihmels *et al.* (2005) (Ihmels et al., 2004) based on detailed analysis of transcriptional regulation in KEGG metabolic pathways. In case of iso-enzymes that are not significantly correlated, Ihmels *et al.* (2005) (Ihmels et al., 2004) found that many of these cases can be explained in terms of differential co-regulation of each iso-enzyme with distinct pathways. Such regulatory structure may help in achieving pathway specific control of fluxes. Our analysis, however, suggests that such conclusions based on concept of pathways should be treated with caution. Indeed, 66% of all reaction in yeast metabolism involves at least one of the highly connected metabolites (which are not completely accounted for in pathways representations). Furthermore, reactions that are catalyzed by iso-enzymes show significantly higher requirement for co-factors (p-value 0.02). Thus the regulation of iso-enzymes can not be simply attributed to pathway structures deduced by neglecting effects of co-factors and other hub metabolites, as several of these metabolites significantly contribute towards organization of transcriptional regulation in metabolism.

7.3.6 Metabolic regulation and dataset used

Analysis presented here is likely to change if more perturbations that are not currently part of this dataset are included in the analysis. Extents of these changes are dependent on how well spread the studied perturbations are over different parts of genome and different environmental conditions. E.g., although NADH was identified as a reporter predominantly in genetic perturbations, it is a top reporter metabolite when environmental perturbation of change from anaerobic to aerobic metabolism is analyzed(Tai et al., 2004) (data not used in this study). However, since structure of metabolic network is independent of the perturbation introduced, metabolite-centered approach will always serve as guideline for understanding regulation in metabolic networks. This is indeed evident from the perturbation/experiment specificity of reporter metabolites observed in this and previous analysis.

7.4 Conclusions

Although metabolic pathways partially exploit the topological connectivity in metabolic networks, they were found to play less significant role in terms of organization of transcriptional regulation than metabolites. Metabolite centered regulation of metabolism is not only limited at transcription level but also extends to metabolome level (Cakir T. *et al.*, Mol Sys Biology, in press, Chapter 5). We argue that regulation of genes surrounding a metabolite is, in part, a thermodynamic and stoichiometric necessity. Thus metabolites provide a robust way to modularize regulation in

metabolic networks. Notably, these metabolic modules are not derived from data, but arise naturally from the network topology. Hence metabolite-based modules are insensitive to underlying data and provide a sound biological explanation to observed regulatory responses. Indeed, mechanistic principles behind such metabolite-centered regulation are being unraveled (Patil K.R. *et al.*, unpublished results, Chapter 6). Same analysis also shows that metabolites provide a basis for rational analysis of emergence of regulatory circuits in metabolic networks. Since, metabolic networks are widely conserved (Peregrin-Alvarez et al., 2003), metabolite-based modules also represent a platform for comparing regulatory architecture across different species.

We further show that it is possible to unravel specific information about regulation at each metabolic node by analyzing whether regulatory response is aimed towards changes in increased/decreased consumption/production of a certain metabolite. Furthermore, through comparison of differential and multi-dimensional response we could deduce whether certain metabolic functions are limited at transcription level or at substrate availability.

An important outcome of this study is the observation that highly connected metabolites contribute significantly towards regulation in metabolism. Thus principles of regulation in metabolism can be completely understood only in an unbiased study where no metabolites are omitted *a priori*. Since these metabolites glue the whole network together, they help cells in rapid adaptation or in tight control of several pathways with minimal changes.

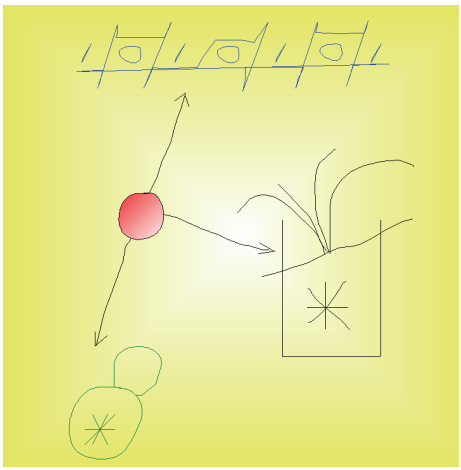
7.5 Supplementary material

Complete results and metabolite scores for differential and multi-dimensional analysis are available at:

http://www2.cmb.dtu.dk/additional_material_for_publications/

Chapter 8: Evolutionary programming as a platform for in silico metabolic engineering

This chapter is based on the publication:
Patil, K. R., Rocha, I., Forster, J. & Nielsen, J. Evolutionary programming as a platform for in silico metabolic engineering. BMC. Bioinformatics 6, 308 (2005).



"Well, in our country," said Alice, still panting a little, "You'd generally get to somewhere else--if you ran very fast for a long time as we've been doing." "A slow sort of country!" said the Queen. "Now, here, you see, it takes all the running you can do, to keep in the same place. If you want to get somewhere else, you must run at least twice as fast as that."

8.1 Abstract

Through genetic engineering it is possible to introduce targeted genetic changes and hereby engineer the metabolism of microbial cells with the objective to obtain desirable phenotypes. However, owing to the complexity of metabolic networks, both in terms of structure and regulation, it is often difficult to predict the effects of genetic modifications on the resulting phenotype. Recently genome-scale metabolic models have been compiled for several different microorganisms where structural and stoichiometric complexity is inherently accounted for. New algorithms are being developed by using genome-scale metabolic models that enable identification of gene knockout strategies for obtaining improved phenotypes. However, the problem of finding optimal gene deletion strategy is combinatorial and consequently the computational time increases exponentially with the size of the problem, and it is therefore interesting to develop new faster algorithms. In this study we report an evolutionary programming based method to rapidly identify gene deletion strategies for optimization of a desired phenotypic objective function. We illustrate the proposed method for two important design parameters in industrial fermentations, one linear and other non-linear, by using a genome-scale model of the yeast *Saccharomyces cerevisiae*. Potential metabolic engineering targets for improved production of succinic acid, glycerol and vanillin are identified and underlying flux changes for the predicted mutants are discussed. We show that evolutionary programming enables solving large gene knockout problems in relatively short computational time. The proposed algorithm also allows the optimization of non-linear objective functions or incorporation of non-linear constraints and additionally provides a family of close to optimal solutions. The identified metabolic engineering strategies suggest that non-intuitive genetic modifications span several different pathways and may be necessary for solving challenging metabolic engineering problems.

8.2 Background

Microorganisms are widely used for producing antibiotics, therapeutic proteins, food and feed ingredients, fuels, vitamins and other chemicals. Currently there is an increasing trend to replace chemical synthesis processes with biotechnological routes based on microbial fermentations. In order to economically produce desired compounds from microbial cell factories it is, however, generally necessary to retrofit the metabolism, since microorganisms are typically evolved for maximizing growth in their natural habitat. Retrofitting of microbial metabolism has traditionally been done through classical strain improvement that involved random mutagenesis and screening, whereas in later years rational design strategies based on genetic engineering have been applied with an increasing success – often referred to as metabolic engineering. In metabolic engi-

neering many experimental and mathematical tools have been developed for introducing directed genetic modifications that will lead to desirable metabolic phenotypes resulting in improved production of desirable compounds or in reduced production of by-products (Nielsen, 2001;Stephanopoulos et al., 1998). Until now most of the successes in metabolic engineering have been based on qualitative or intuitive design principles. However, even though there are several success stories in metabolic engineering there are also many attempts that have failed due to the lack of rational strategies based on predictive analysis tools.

Microbial metabolism is often subjected to tight regulation and is constrained by mass and energy conservation laws on a large number of intracellular metabolites, and this makes it difficult to predict the effects of introducing genetic modifications in a given cell. Moreover, as metabolic pathways and related regulatory processes form complex molecular and functional interaction networks (Patil and Nielsen, 2005;Ideker et al., 2001), it is only through analysis of the metabolism as a whole in an integrative systems approach (Stephanopoulos et al., 2004) that one may evaluate the effect of specific genetic modifications. Genome-scale models of microbial organisms (Price et al., 2003), comprising different levels of information, primarily on the stoichiometry of the many different reactions but possibly also comprising some information about regulation, could offer a suitable platform for developing systems level tools for analyzing and engineering metabolism (Patil et al., 2004) (Chapter 2). Although there have been some attempts to simulate dynamic behavior of whole cell systems (Tomita, 2001;Covert et al., 2004), currently these approaches enjoy limited applicability due to lack of kinetic and regulatory information on the whole genome-scale. Nevertheless, in absence of kinetic and regulatory information it is possible to at least partly predict the behavior of cellular metabolism by using steady state analysis based on genome-scale stoichiometric models.

Genome-scale stoichiometric models represent the integrated metabolic potential of a microorganism by defining flux-balance constraints that characterizes all feasible metabolic phenotypes under steady state conditions. Because of the large number of reactions occurring in cellular metabolism, the dimensions of the solution space (or the number of feasible metabolic phenotypes) defined by genome-scale models (Schilling et al., 2000;Schuster et al., 2000) is very large. Consequently, combinatorial complexity prevents calculation of all feasible metabolic phenotypes that a microbial genotype can assume under a given environmental conditions (Klamt and Stelling, 2002). One of the approaches to determine the metabolic phenotype (i.e. the fluxes through all metabolic reactions) is to use flux balance analysis (FBA) (Kauffman et al., 2003;Fell and Small, 1986). In FBA a particular flux or a linear combination of various fluxes (objective function) in

the model is optimized through linear programming, thus leading to a solution to the fluxes through all metabolic reactions. Since several microbial metabolic networks have evolved towards operation of optimal growth rate (Ibarra et al., 2002;Edwards et al., 2001;Burgard and Maranas, 2003;Famili et al., 2003), the use optimization of growth rate is an often applied objective function in FBA. There are, however, some other approaches to determine flux distributions, especially for deletion mutants that might not be capable of realizing the same objective function as the wild-type strain (Segre et al., 2002;Shlomi et al., 2005;Beard et al., 2002). Nevertheless, all these methods provide a basis for using genome-scale metabolic models to predict possible metabolic phenotypes, and hence for *in silico* metabolic engineering. However, despite of their potential, genome-scale stoichiometric models have been scarcely used for metabolic engineering purposes.

The algorithm developed by Maranas *et al.* (Burgard et al., 2003;Pharkya et al., 2004) (named OptKnock) represents one of the first rational modeling frameworks for suggesting gene knockouts leading to the overproduction of a desired metabolite. OptKnock searches for a set of gene (reaction) deletions that maximizes the flux towards a desired product, while the internal flux distribution is still operated such that growth (or another biological objective) is optimized. Thus the identified gene deletions will force the microorganism to produce the desired product in order to achieve maximal growth. Indeed, the design philosophy underlying OptKnock approach takes advantage of inherent properties of microbial metabolism to drive the optimization of the desired metabolic phenotype. The relation of OptKnock with the biological objectives of microorganisms makes it an attractive and promising modeling framework for *in silico* metabolic engineering.

OptKnock is implemented by formulating a bi-level linear optimization problem using mixed integer linear programming (MILP) (Burgard et al., 2003) that guarantees to find the global optimal solution. In this report, we extend the applicability of OptKnock approach by formulating the *in silico* design problem by using a Genetic Algorithm (GA), hereafter referred to as OptGene. Genetic algorithms use the principle of Darwinian evolution to search (*evolve* through mutations and reproduction) for the global optimal solution (individual with a maximum *fitness* score). Direct relation of GA with biological evolution makes it a natural method of choice to identify suitable genetic modifications for improved metabolic phenotype. There are two major advantages of the OptGene formulation. Firstly, OptGene demands relatively less computational time and thus it enables to solve problems of larger size. This is of particular importance as the relation between the size of the problem (as defined by the number of enzymes and number of deletions

desired) and the corresponding search space (combinations of enzymes to be deleted) is combinatorial (Supplementary Figure S-8.1). Thus, the number of possible combinations of 5 reaction-deletions in a model with 250 reactions is more than 7.8×10^{9}, whereas existing genome-scale stoichiometric models comprise a significantly higher number of reactions. Secondly, the OptGene formulation allows the optimization of non-linear objective functions, which is of considerable interest in several problems of commercial interest. One example of an important non-linear engineering objective function is the productivity (amount of product formed per unit time).

8.3 Results and Discussion

8.3.1 OptGene algorithm

Two different versions of the OptGene algorithm were used in this work, differing mainly on the representation of the metabolic genotype: binary (binOptGene) and integer (intOptGene) representations. The binary form of the OptGene algorithm is schematically illustrated in Figure 8.1, and the important steps of both representations are explained in the following.

8.3.2 Model pre-processing

Since GA do not exhaustively search the complete solution space, it is necessary to avoid local optimal solutions by proper formulation of the problem. We therefore pre-processed the model to remove duplicate and dead-end reactions. Also a linear pathway (or enzyme subset (Pfeiffer et al., 1999)) was represented as a single reaction in GA. Moreover, lethal reactions (including genes that are found to be lethal *in vivo*, but not *in silico*) were not included as the possible targets in GA. This pre-processing step reduced the problem size considerably and thus reduced the number of local optimal solutions (data not shown).

8.3.3 Chromosome representation of metabolic genotype

Each reaction in the metabolic model can be associated with one or more genes in the genome. In the binOptGene algorithm each of those genes is represented by a binary variable indicating its absence/presence (0/1), and thus a set of these variables forms an "individual" (sometimes also referred to as "chromosome" in evolutionary algorithms nomenclature) representing a particular mutant that lacks some metabolic reactions when compared with the wild type (Figure 8.2). For the intOptGene implementation, the individuals are composed of integer numbers representing only the genes to be deleted, according to their relative order in the metabolic model. This way, the number of genes to be deleted can be directly imposed by changing the size of the individuals. The phenotypes of every individual can be obtained by using FBA or other algo-

rithms. The problem then is to find the set of genes to be deleted from an individual so as to obtain a desired phenotype (e.g. with maximum product yield and minimum undesired by-product yield).

8.3.4 Initialization of population

The GA begins with a predefined number of individuals, forming a population. In the binOptGene, individuals in the population can be initialized in different ways, e.g. by assigning present/absent status to each gene randomly, or assigning present status to all genes, while in the intOptGene representation, the population is usually initialized randomly.

8.3.5 Scoring fitness of individuals

Each individual is assigned a fitness score that determines whether it will reproduce and/or propagate to the next generation. The fitness score of an individual is calculated using the desired objective function value. The objective function value can be calculated using FBA, minimization of metabolic adjustment (MOMA) (Segre et al., 2002), regulatory on-off minimization (ROOM) (Shlomi et al., 2005) or any other algorithm. The GA by itself is independent of scoring algorithm.

8.3.6 Crossover of chromosomes

After the fitness score is calculated for all individuals in the population, the best individuals are selected for crossover. A selection scheme that is most commonly used is the Roulette wheel, where individuals are selected based on the magnitude of the fitness score relative to the rest of the population. The higher the score, more likely an individual will be selected. Selected individuals are then crossed to produce a new offspring. The crossover methods used in this study were one-point, two-points, and uniform crossover (Goldberg, 1989).

8.3.7 Mutation

Individuals propagating to the new population are mutated (in our formulation, a gene is deleted) with a given probability.

8.3.8 New population and termination

Mutation and crossover give rise to a new population, which can then again be subjected to a new round of evaluation, crossover and mutations. This cycle is repeated until an individual with a satisfactory phenotype is found.

We illustrate the principles and utility of OptGene algorithm by using three interesting metabolic engineering problems with the yeast *Saccharomyces cerevisiae*, which is one of the most widely used cell-factories. We applied OptGene for *S. cerevisiae* to identify gene-deletion strategies for improving yield and substrate-specific productivity of three metabolites, namely vanillin, glycerol and succinate. The yield of a product (metabolite) of interest is defined as the grams of product produced per unit gram of the substrate consumed, whereas substrate-specific productivity is defined as the grams of product produced per unit time per unit substrate consumed. It is important to note that models based only on stoichiometry can not predict rates without an assumption of a fixed substrate uptake rate. Since the substrate uptake rates for deletion mutants might change substantially and the fact that it is very difficult to predict such changes *a priori*, in general the productivity can not be optimized by using stoichiometric models. One of the ways to circumvent this problem is to optimize the function [Product Yield × Growth]. Although, this quantity will be equal to the substrate-specific productivity under the assumption of a fixed substrate uptake rate, we will refer to this term as Biomass-Product Coupled Yield (BPCY) rather than the productivity as this may cause confusion (also see Note 1 for comments about the growth rates for mutants). BPCY represents an interesting example of a non-linear objective function that can be optimized by using OptGene.

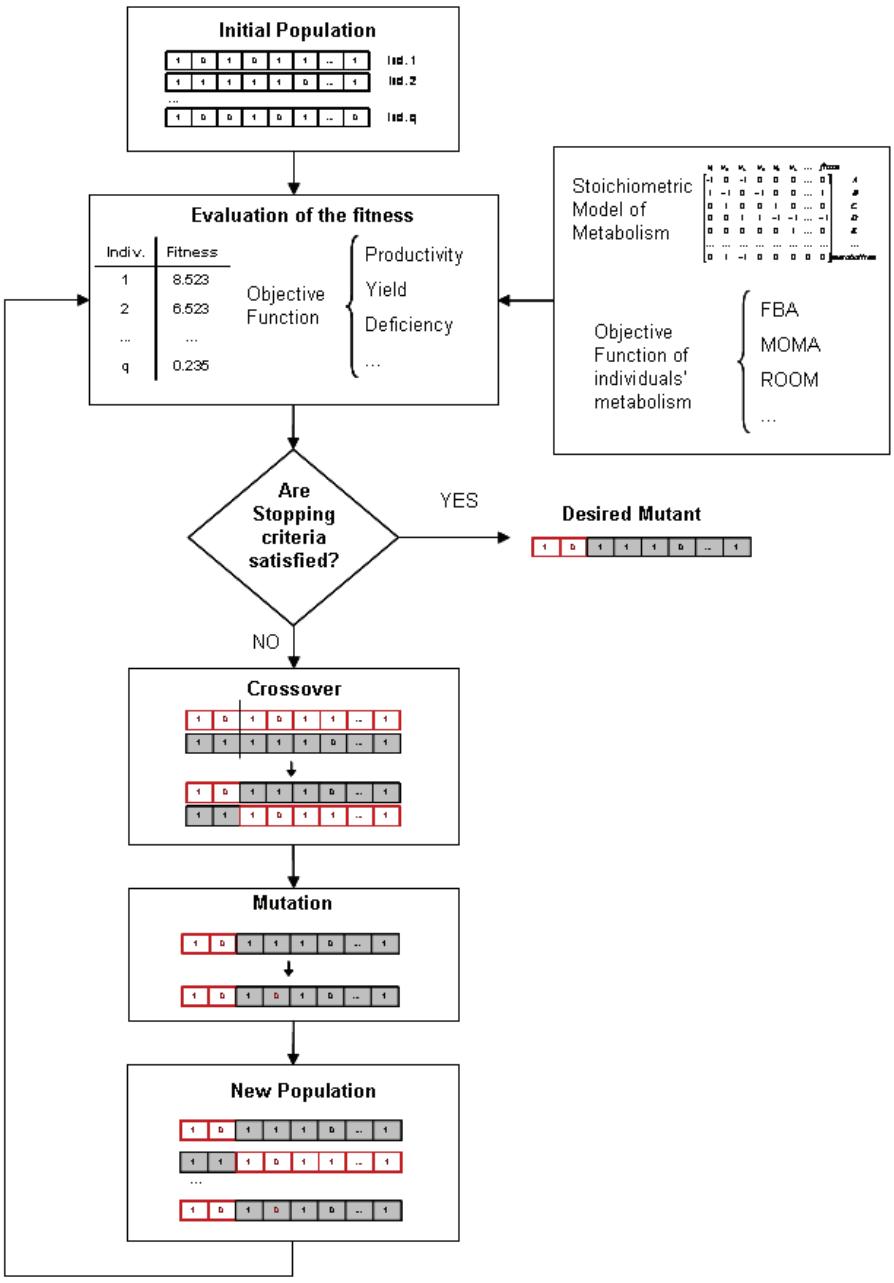


Figure 8.1. Schematic overview of the OptGene algorithm. A population of individuals is initiated by specifying a present/absent status for each gene in each of the individuals. Individuals are then scored for their fitness by using FBA/MOMA/other method of choice and the objective function (/s). Individuals are selected for mating based on their fitness score, and subsequently crossed to produce new offspring. Mutations are introduced in individuals randomly at specified mutation rate and thus a new population is obtained. This cycle of evolution is repeated until a mutant (or mutants) with a desired phenotypic characteristics is obtained. Please refer to the text for detailed description of each step in the algorithm. Grey shaded or red walled boxes are used to represent different individuals in the cross-over process.

Ind.- Individual. FBA- Flux balance analysis (Fell and Small, 1986;Kauffman et al., 2003). MOMA- Minimization of the metabolic adjustment (Segre et al., 2002). ROOM- Regulatory on/off minimization (Shlomi et al., 2005).

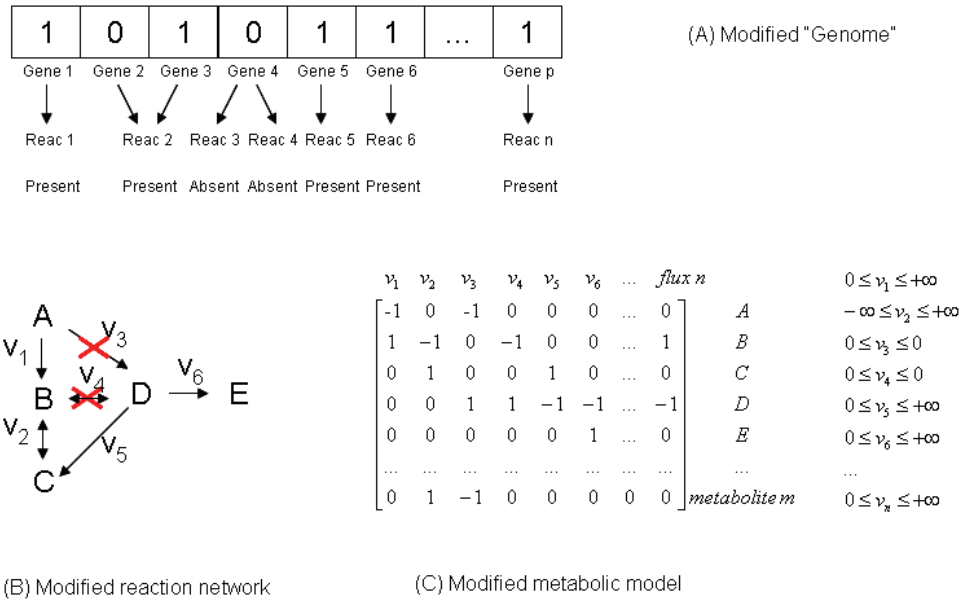


Figure 8.2. Representation of the metabolic genotype. Each gene of the microorganism is assigned a binary value, representing its absence/presence in the mutant (A). The individual genes are associated with one or more reactions in the metabolic network (B). When a given reaction is in the absent status, the upper and lower bonds for the corresponding metabolic flux are set to zero, resulting in a modified metabolic model (C).

8.4 Vanillin case study

Vanillin is a natural flavor compound extracted from plants and is widely used as a food ingredient. There is some commercial interest in producing vanillin by using recombinant microorganisms and in particular *Saccharomyces cerevisiae* which is a food grade organism. Since vanillin is not produced naturally by *S. cerevisiae*, the corresponding reactions were inserted into the model as suggested by Pharkya *et al.*(Pharkya et al., 2004). Then we used OptGene to find gene deletion strategies to improve the BPCY as well as the yield of vanillin. We found that it was possible to improve the vanillin yield *in silico* up to 90 % of the theoretical limit by deleting only 2 reactions (pyruvate decarboxylase and glutamate dehydrogenase), while keeping the growth rate at 60% of the parental strain. A similar strategy was predicted for a mutant with the maximum BPCY. The suggested strategy diverts the pyruvate flux going to ethanol towards vanillin where NADH is oxidised back to NAD^+. Furthermore, deletion of glutamate dehydrogenase results in an increased availability of NADPH needed for vanillin biosynthesis. Increasing the allowable number of deletions did not result in substantial improvement in the yield or BPCY.

8.5 Glycerol case study

Currently glycerol is mainly recovered as a by-product from soap manufacturing or produced from propylene and is widely used to synthesize several products ranging from cosmetics to lubricants (Wang et al., 2001). Alternatively, glycerol can also be produced through microbial fermentation using sustainable carbohydrate resources. *Saccharomyces cerevisiae* naturally produces glycerol in small quantities during anaerobic fermentation or under osmotic stress. Moreover, glycerol plays an important role in maintaining the cytosolic redox balance under anaerobic conditions and it is therefore interesting to study the effects of gene-deletions on yield and productivity of glycerol. We applied the OptGene algorithm to identify gene deletions leading to improved yield and BPCY of glycerol under aerobic conditions, where the maximum theoretical yield of glycerol is much higher as opposed to anaerobic fermentation.

Results suggested that no single gene deletion will result in glycerol production, whereas a strategy for double reaction deletion is identified, namely FBP1 (Fructose-1,6-bisphosphatase) and genes encoding Glyceraldehyde-3-phosphate dehydrogenase (*TDH1*, *TDH2* and *TDH3*). This strategy makes intuitive sense as reactions that branch the flux away from dihydroxyacetone phosphate (the precursor for glycerol) are deleted (see figure 8.3 for a schematic representation of yeast central carbon metabolism). With this strategy it is possible to obtain a yield of 0.49 g/g-glucose with a corresponding growth rate that is 80% lower than the reference strain. Increasing the number of deletions up to six did not result in a further substantial increase in the yield. However, interestingly, the BPCY of glycerol improved with the number of deletions allowed. With six deletions, the BPCY reached up to 41 mg/g glucose-hr (yield of 0.31 g/g-glucose) with a growth rate equal to 50% of that of the reference strain. Moreover, the identified deletions for yield and BPCY improvement are different (Supplementary Table S-8.2). Notably, the suggested deletions span not only the central carbon metabolism but also extend to amino acid and vitamin metabolism, illustrating the tight links between different metabolic pathways arising from the mass balance constraints. This also illustrates the need for the here reported algorithm which can search this vast solution space efficiently.

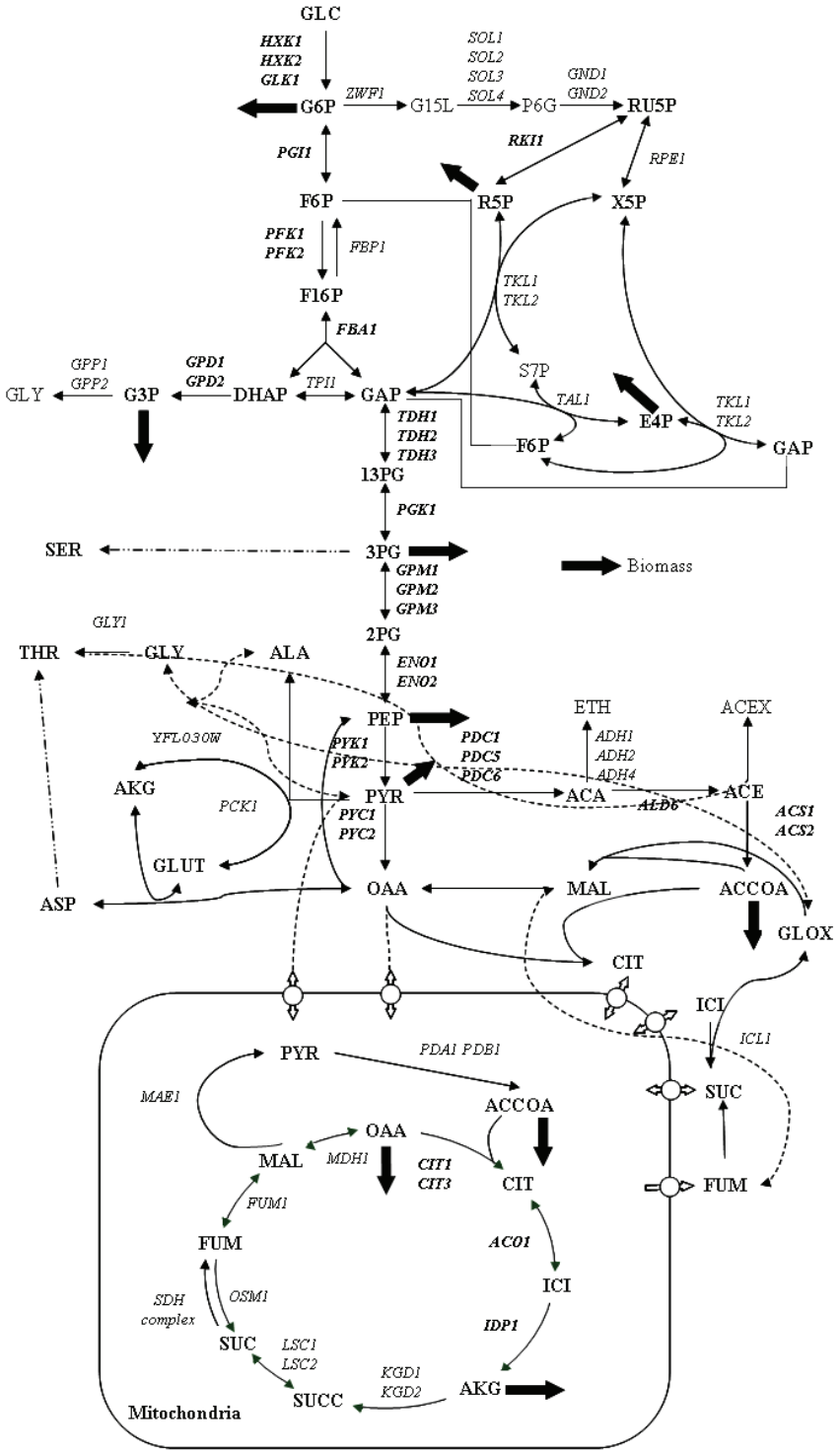


Figure 8.3. Schematic overview of the *Saccharomyces cerevisiae* central carbon metabolism. The figure shows important pathways in the central carbon metabolism including certain branch points towards the amino acid metabolism. The thick arrows indicate the drain of metabolites towards biomass production. Arrows with the style indicates a lumped pathway. Multiple names for a reaction indicate the presence of iso-enzymes. The nomenclature of the metabolites can be found in the Supplementary table S-8.1. The figure is partially adapted from Forster et al. (2002) (Forster et al., 2002).

8.6 Succinic acid case study

Succinic acid is one of the intermediates of the TCA cycle and is an interesting chemical to be used as a feedstock for synthesis of a wide range of chemicals. As a metabolite from the central carbon metabolism, it is a good case study for devising metabolic engineering strategies. Multiple gene deletion strategies obtained using OptGene algorithm for improving succinic acid yield and BPCY are summarized in table 8.1.

Firstly, we note that the maximum theoretical yield of succinic acid is 0.506 g/g glucose (Note 2) when no biomass is being produced, and that no succinic acid can be produced at optimal biomass growth rate. Moreover, no single gene deletion strategy resulted in succinic acid production. For a double gene deletion strategy, deletion of the *SDH-complex* (succinate dehydrogenase) and *THR1* (homoserine kinase) is predicted to result in a succinic acid yield of 0.018 g/g glucose, with a 10% reduction in the growth rate. Flux re-distribution leading to this improvement in the double-deletion mutant is quite interesting and non-intuitive. Deletion or inactivation of the *SDH-complex* prevents the conversion of mitochondrial succinate to fumarate, while simultaneous deletion of *THR1* forces threonine synthesis via glycine, which may be formed from glyoxylate. Consequently there is increased flux through *ICL1* (cytosolic isocitrate lyase, catalyzing the reaction from isocitrate to glyoxylate and succinate), thus creating surplus succinate that is secreted by the cell. Moreover, this flux re-distribution is also associated with an increased flux through the pentose phosphate (PP) pathway for increased NADPH availability. We note that in the mutant with only the *SDH-complex* deleted, threonine is synthesized via aspartate, which is optimal route for maximizing biomass production. The same double deletion mutant was also predicted to show maximum BPCY (4.5 mg/g glucose-hr).

The search for a triple deletion mutant with maximum succinate yield suggested deletion of the *SDH-complex*, *ZWF1* (Glucose-6-phosphate dehydrogenase) and *PFK2* (Phosphofructokinase). Although this resulted in increased prediction of succinate yield (0.21 g/g glucose), the corresponding growth rate is very low (96% reduction in growth rate), making this solution unattractive. However, a triple deletion mutant with maximum BPCY (16 mg/g glucose-hr) was found to have 76 % of the wild-type growth rate and a succinate yield of 0.07 g/g glucose. The corresponding solution suggested deletion of *SER3* in addition to the double deletion strategy discussed above. Deletion of *SER3* blocks the synthesis of L-Serine via 3-Phospho-D-glycerate, which increases the demand on glycine production via glyoxylate. Overall, it leads to a further increase in the flux through *ICL1* ensuring a higher flux towards succinate while maintaining a

high growth rate. This increase is also associated with a further increase in the flux through the PP pathway.

In spite of a slow growing triple deletion mutant for improved yield, the algorithm found a quadruple deletion mutant with not only improved yield (0.36 g/ g glucose), but also with much higher growth rate, as compared to the triple deletion mutant (table 8.1), and therefore higher BPCY. The suggested genes for deletion are the *SDH-complex*, *ZWF1*, *PDC6* (pyruvate decarboxylase) and *AGP3* (glutamate permease). Deletion of *ZWF1* increases the flux through glycolysis and deletion of *PDC6* increases conversion of pyruvate to oxaloacetic acid via *PYC1*. This flux could be directed towards glutamate production and into the TCA cycle. But since the *SDH-complex* is deleted the flux through TCA cycle is limited, while deletion of secretion reaction for surplus glutamate forces malate formation from oxaloacetic acid. The flux through malate is then directed to succinate via fumarate. We also searched for a quadruple deletion strategy for maximum BPCY and the algorithm suggested the same deletion strategy as for the maximum yield, with a corresponding BPCY of 29 mg/g glucose-hr. This BPCY shows a substantial increase over the BPCY obtained with the triple deletion strategy.

Results of a further search allowing more gene deletions, for improvement in yield and BPCY, are summarized in table 8.1. Here we note that it might be difficult to realize some of the suggested optimal strategies *in vivo* due to a variety of reasons, e.g. regulatory constraints, orphan reactions etc. However OptGene provides not only the optimal solution found, but also generates a family of "good" solutions and thus provides many strategies that can be further analyzed manually before experimental verification. Some of such alternative solutions are also reported in table 8.1.

Table 8.1. Different deletion strategies suggested by OptGene algorithm for improving succinate yield and Biomass Product Coupled Yield.

Objective function	Number of deletions	Suggested deletions[1]	Objective function value[2]	%Maximum Growth	Unique solution?[3]
Succinate yield	5	SDH-complex, ZWF1, PDC6, U133, U221	0.39	14%	Yes
		SDH-complex, ZWF1, PDC6, U133, U41	0.37	1%	Yes
	4	SDH-complex, ZWF1, PDC6, AGP3	0.356	30%	Yes
	3	SDH-complex, ZWF1, PFK2	0.211	4%	Yes
		SDH-complex, SER3, THR1	0.074	76%	Yes
Succinate Biomass Product Coupled Yield	4	SDH-complex, ZWF1, PDC6, AGP3	29	30%	Yes
		SDH-complex, SER3, THR1, U221	22	75%	Yes
	3	SDH-complex, SER3, THR1	16	76%	Yes
		SDH-complex, ZWF1, GLT1	9.78	42%	Yes

[1] Only few of the suggested strategies, with high objective function values are shown. OptGene found many strategies with different, but high objective function values. This tendency can be controlled by varying GA parameters.
[2] Units are: Yield in gram (gram glucose)$^{-1}$,
Biomass Product Coupled Yield in milli-gram (gram-glucose.hour)$^{-1}$
[3] Uniqueness of the solution was verified by first optimizing for the biomass, and then minimizing and maximizing the succinate flux at fixed, optimal biomass value.

8.7 MOMA approach

The examples discussed above use FBA as scoring function to evaluate fitness of an individual in the GA. However, as noted before, the flux distribution of mutants of *Echerichia coli* have been shown to be better approximated by assuming that the fluxes tend to have a minimum distance from wild-type flux distribution, which may not correspond to the flux distribution for maximum

growth (Segre et al., 2002). Nevertheless, although this approach, referred to as Minimization of Metabolic Adjustments (MOMA), might explain the flux distribution of mutants better than FBA, such mutants might approach towards FBA-predicted optimal solution when evolved under growth pressure (Fong et al., 2003;Ibarra et al., 2002).

To check whether the two approaches for evaluating flux distributions (namely FBA and MOMA) result in different predictions for multiple deletion mutants, we used OptGene to search for double and triple deletion mutants with improved succinic acid yield and BPCY. The double deletion strategy for obtaining maximum yield with MOMA includes deletion of *FUM1* (fumarase) and *PDA1* (pyruvate dehydrogenase). This strategy is different from that suggested by using FBA, and it also predicts a better yield (0.11 g/g glucose) for a double deletion mutant. In case of BPCY the MOMA approach yielded the same productivity, although with different genes (*RPE1* and an orphan reaction in mitochondria). However, an effective comparison of FBA and MOMA for multiple deletion mutants can only be done after experimental evaluation.

8.8 Significance and effects of different GA parameters

Parameterization of stochastic optimization methods like evolutionary algorithms is recognized as a difficult task and for this particular problem only an empirical study of the effect of different parameters was conducted. The main purpose of this parameterization was to be able to obtain a global optimum within a reasonable computation time.

Different sizes of the population were tested, and it was found that an increase beyond 125 individuals did not improve the results significantly. Furthermore, a mutation rate of 1/(genome size) was found to be optimal for both representations (Supplementary Figure S-8.2) .

Regarding crossover methods, for the binOptGene representation, one-point, two-points and uniform crossover methods were tested, and the different crossover techniques gave almost the same results, indicating that all approaches are equally good, probably due to their similar operation mode. For intOptGene, only one type of crossover method was tested, namely uniform crossover where a child obtains a gene from each parent with equal probability.

After parameterization, for both representations, and for a typical optimization run, the evolutionary algorithms were able to achieve a solution within 1000 generation, although the algorithm was allowed to run until 5000 generations. A typical convergence curve can be found in Figure 8.4.

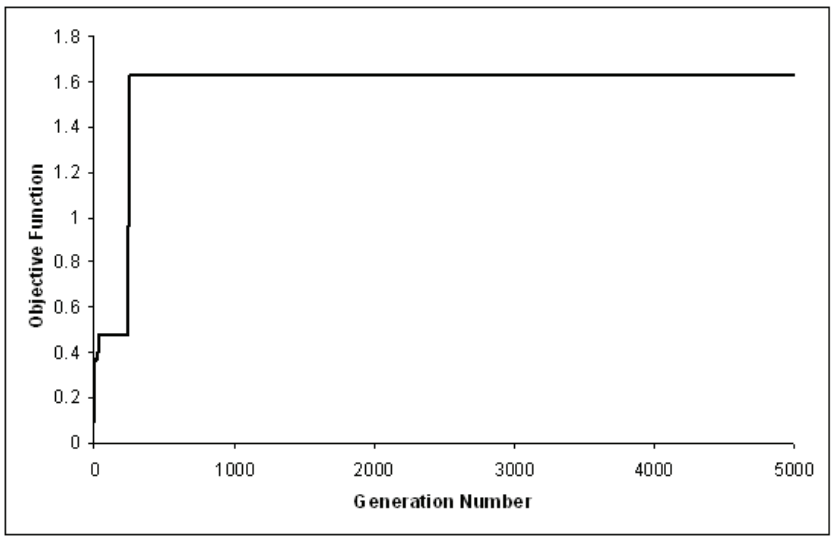


Figure 8.4. Typical shape of the convergence curve of OptGene.

8.9 Resemblance to Natural Evolution

The theoretical foundations of genetic algorithms rely on a notion of short, highly fit schemata, also known as building blocks (see e. g. (Michalewicz, 1996;Goldberg, 1989)), that are propagated generation to generation and constitute the basis for the convergence to optimal solutions. For the strain design problem, building blocks can be regarded as subsets of genes in a close position on the individuals of the evolutionary algorithms that, when deleted together, improve process yield or productivity.

The differences on the representation of individuals in both approaches used in this work originate different requisites in terms of the formation of building blocks: as in the binary representation the order of the genes in the individuals follows closely that of the stoichiometric model (where genes are grouped according to the main pathways they integrate), only related genes can be a part of the building blocks. On the other hand, for the integer representation any subset of unrelated genes can form a building block. A natural conclusion of this observation is that the more the genes in the metabolic model follow a biological meaningful order, the more similar the binOptGene optimization approach is to a biological evolution of microorganisms under a given selective pressure.

Additionally, we observed that if the limitation on the number of genes to be deleted in binOptGene is imposed by using penalty functions after evaluation of individuals, the number of invalid individuals in the population at a given generation is very large and consequently negatively affects the convergence.

Nevertheless, in spite of the described differences, and although it is known that usually Genetic Algorithms do not perform very good for problems of the size found for the binary implementation, similar results were obtained for both approaches, after parameterization. In fact, for the majority of the runs, and with both representations, there was a clear convergence to an optimum (Figure 8.4), and the solutions found were very similar among all the repetitions (typical values of the relative standard deviation of 20 runs are 6%). Additionally, most of the times the final solution was found very early, indicating that 500-1000 generations are probably enough for converging to a satisfactory solution. However, by looking at the shape of the convergence curve in figure 8.4, it is clear that there are several sudden increases in the performance of the best individual, as opposed to the most often observed smooth convergence curves obtained with evolutionary algorithms. These step changes in the objective function value are usually an indication that the optimization is being stopped very prematurely but, as more iterations do not improve the final solution, it is more likely that the problem itself is discrete. In fact, and although additional characterization of the search space is needed, this observation can be explained by the evidence that, when a good candidate for deletion is found, the performance of the best individual in a population increases significantly.

8.10 Global optimal solution and computational cost

In case of succinate yield optimization, the optimality of the solution found by OptGene was verified using exhaustive search with up to 4 deletions. In case of BPCY, although the optimal solution for 3 deletions represented a global optimum, for a 4 deletion case OptGene found a sub-optimal solution. However, this solution was quite close to the global optimal solution (85% of the global optimal value). With 5 deletions the optimal solution found reached quite close to the maximum possible BPCY value. We hereby note that in cases where global optimality can not be directly verified, a good estimate for closeness to the global optimal solution can be found by using a curve similar to that presented in Supplementary Figure S-8.3. The plot in the Supplementary Figure S-8.3 is generated by fixing the biomass yield at different values and then optimizing for the succinate production.

The computational cost of OptGene (estimated as the number of objective function evaluations necessary to find an optimal solution) was found to be 0.03 % of that found by using exhaustive search for 4 gene deletion case (succinate yield) and 0.33 % for succinate BPCY case. However, we did not observe any direct correlation between the number of deletions and the computational cost. Supplementary table S-8.4 summarizes the computational cost associated with the succinic acid optimization case.

8.11 Multiple optima

Since the flux distribution obtained using FBA is not necessarily unique, the objective function value obtained in the fitness evaluation routine may not be unique as well. This is usually due to the possibility of other by-products being formed instead of the desired product (Supplementary Table S-8.3 lists the metabolites that were allowed to be secreted by the cells in this study). Consequently it has an important implication while designing the deletion strategies. Such check for uniqueness of objective function can easily be incorporated in the fitness evaluation routine by using flux variability analysis. Thus, e.g., an upper and lower bound can be calculated for the product flux at the optimal growth rate. The choice between "pessimistic" and "optimistic" fitness value can be left for the user. However, we note that for the results presented in this study, the solutions obtained were unique as indicated in the last column of Table 8.1.

8.12 Conclusions

We report a GA based framework termed OptGene for designing microbial strains *in silico*. OptGene presents two major advantages, higher speed and ability to optimize for non-linear objective functions. The optimal solution for a four deletion problem (succinate yield case) was found using OptGene by searching only 0.03% of the total solution space. For a higher number of deletions, the OptGene search space represents considerably lower fraction of the total solution space that increases exponentially. As a consequence of an exponential increase in the search space, a detailed study of the correlation between the OptGene search space and the total solution space was not feasible. Nevertheless, as discussed in the results section, it is possible to estimate the closeness to the global optimal solution by comparing the results with the plots as reported in Supplementary Figure S-8.3 . Consequently, high computational speed of OptGene enables addressing the problems involving large number of genes, and searching for higher number of deletions. This is of particular interest as genome-scale models of simple eukaryotic organisms like *S. cerevisiae* include more than 1000 reactions. In case of simple minimal media that we used in our simulations, this set of 1000 reactions can be reduced to 240 reactions as described in the algorithm. This number can still be large for solving quadruple deletion problem using exhaustive search algorithms.

The metabolic engineering strategies reported in this work suggest that non-intuitive genetic modifications spanning several different pathways may be necessary for solving challenging metabolic engineering problems. Consequently *a priori* selection of candidate targets might lead to sub-optimal solution, and it is desirable to consider the whole model. Moreover, with the recent advances on the experimental front, it is feasible to construct mutants with many knockouts in

real time. It should also be noted that we might often need to recalculate the results in case of changes/errors in the model, e.g. after including regulatory information or addition of a new reactions. Speed of calculations can be a key factor in such cases. OptGene can serve to provide a quick hint to whether a particular function of interest can be improved at all or up to what extent. The ability of OptGene to optimize for non-linear objective functions opens new opportunities for designing microbial strains with tailor-made metabolic phenotype, e.g. a strain with high BPCY of x and low yield of y.

The GA formulation can provide us with multiple solutions, and thus an opportunity to choose from many good solutions. This is of interest as many of the predicted solutions might be difficult to realize due to complex biological regulation, which is difficult to account for in scoring function models. Moreover, the GA framework is very flexible and thus can easily be changed to use different scoring functions depending on the problem and system under investigation. In conclusion, OptGene represents a computationally efficient, flexible and natural tool for *in silico* designing of microbial strains by using genome-scale models.

8.13 Methods

8.13.1 Metabolic model

Genome-scale reconstruction of *S. cerevisiae* reported by Förster et al. (Forster et al., 2003a) was used as stoichiometric model of yeast metabolism. All simulations were performed for aerobic glucose-limited conditions. The glucose uptake rate was fixed to 3 mmoles/gDW/hour while the maximum oxygen uptake rate was set to 9 mmoles/gDW/hour (Overkamp et al., 2000).

8.13.2 FBA and MOMA

FBA simulations were performed using the GNU linear programming kit (http://www.gnu.org/software/glpk/glpk.html), while MOMA calculations were performed by using an Object oriented quadratic programming package (Gertz and Wright, 2003).

8.13.3 Genetic algorithm

The genetic algorithm was implemented as a C++ program using the GAlib package (http://lancet.mit.edu/ga/).

Note 1: Reported growth rates for mutants

As discussed in the main text, FBA (and other steady state models) can not simulate "rate" without specification of the specific substrate uptake rates (substrate uptake per unit biomass per unit

time). Consequently the reported growth rates for the mutants should be more correctly interpreted as biomass yields.

Note 2: Maximum theoretical yield of succinate

The maximum theoretical yield of succinic acid reported in this study is calculated using FBA, whereas external H^+ was balanced. In case where H^+ is regarded as unbalanced (or external) metabolite, maximum yield is 0.98 g/g glucose. This difference is very high and hence can result in big differences in the predictions reported. However, the choice is not trivial since the exact mechanism by which succinic acid is transported out of cell is unknown. Moreover, in case where H^+ is not balanced, certain contradictions with the experimental observations were found under anaerobic conditions. For this reason we chose to use a conservative estimate for the maximum theoretical yield. We also note that the theoretical yields were calculated with the constraints for maintenance cost, and no CO_2 uptake. Thus the reported yields are slightly lower than the stoichiometric yields (1.124 g/g glucose in case of succinate).

Note 3: Data availability

The flux distributions, model reactions and other data related to this article can be obtained for non-profit research purposes by contacting the corresponding author (JN).

Acknowledgements: Authors are grateful to Miguel Rocha, Ana Paula Oliveira, John Villadsen and Donatella Cimini for fruitful discussions. IR is grateful for the financial support provided by FCT (Portuguese Science Foundation) under the Post-Doctoral grant BPD 11634/2002.

8.14 Supplementary material

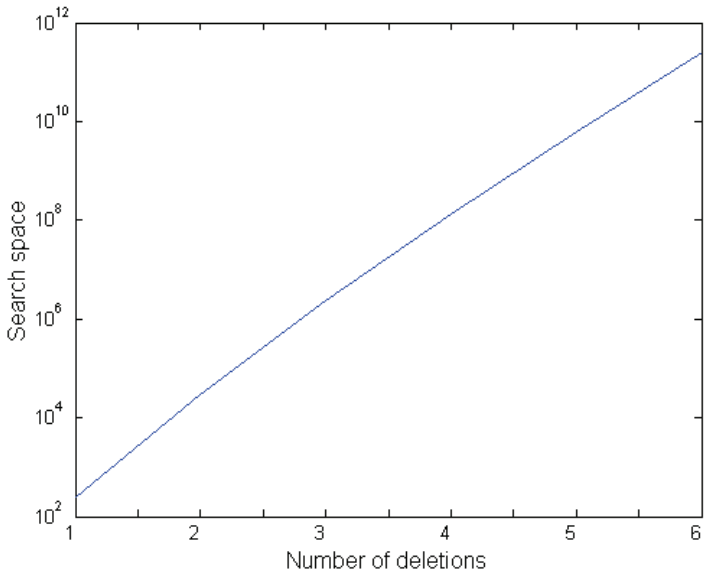


Figure S-8.1. Increase in size of search space (number of possible combinations for given number of deletions) for a reaction-deletion problem with 240 reactions.

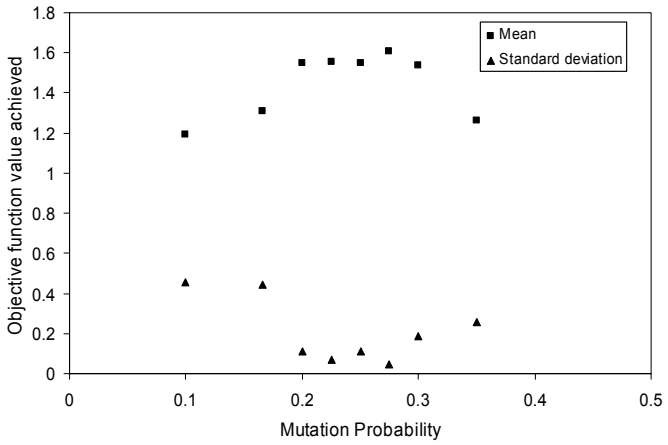


Figure S-8.2. Effect of mutation probability on performance of OptGene. Mean and standard deviation for Succinic acid yield for best individual after fixed number of generations are plotted (arbitrary units).

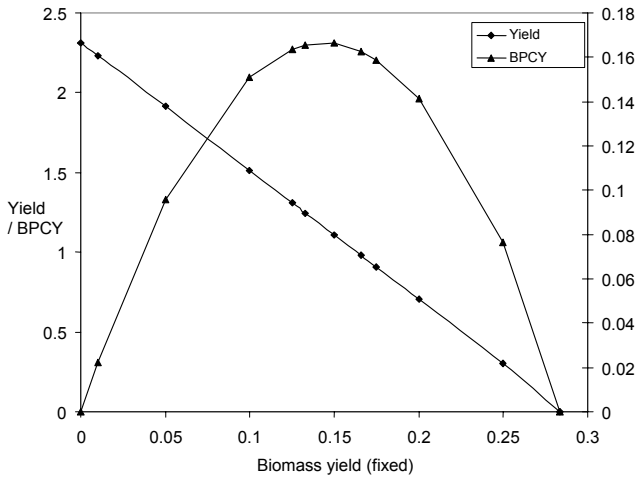


Figure S-8.3. Maximum yield and Biomass Product Coupled Yield of succinate at fixed biomass yield and glucose uptake rate (arbitrary units).

Table S-8.1. Metabolite abbreviations used in the Figure 3 of main text.

Abbreviation	Metabolite
ACCOAcyt	cytosolic Acetyl Coenzyme-A
ACCOAmit	mitochondrial Acetyl Coenzyme-A
ACA	Acetaldehyde
ACE	Acetate
ACEX	Acetate (extracellular)
AKG	2-Oxoglutarate
CIT	Citrate
CO2	CO2
DHAP	Dihydroxyacetone phosphate
ETH	Ethanol
E4P	D-Erythrose 4-phosphate
F6P	D-Fructose 6-phosphate
F16P	D-Fructose 1,6-bisphosphate
FUM	Fumarate
GA3P	D-Glyceraldehyde 3-phosphate
G6P	D-Glucose 6-phosphate
G15L	D-Glucono-1,5-lactone 6-phosphate
GLOX	Glyoxylate
GLC	Glucose
GP	sn-Glycerol 3-phosphate
ICI	Isocitrate
MAL	Malate
OAA	Oxaloacetate
P13G	3-Phospho-D-glyceroyl phosphate
P2G	2-Phospho-D-glycerate
P6G	6-Phospho-D-gluconate
P3G	3-Phospho-D-glycerate
PEP	Phosphoenolpyruvate
PYR	Pyruvate
R5P	D-Ribose 5-phosphate
RU5P	D-Ribulose 5-phosphate
SUC	Succinate
SUCCOA	Succinyl-CoA
S7P	Sedoheptulose 7-phosphate
X5P	D-Xylose-5-phosphate
ADP	ADP
ATP	ATP
NADPcyt	cytosolic NADP+
NADHcyt	cytosolic NADH
NADcyt	cytosolic NAD+
NADPHcyt	cytosolic NADPH
NADmit	mitochondrial NAD+
NADHmit	mitochondrial NADH
NADPmit	mitochondrial NAD+
NADPHmit	mitochondrial NADPH
FAD	FAD++
FADH2	FADH2
SER	Serine
THR	Threonine
ASP	Aspartate
GLUT	Glutamate

Table S-8.2. Different deletion strategies suggested by OptGene algorithm for improving glycerol yield and Biomass Product Coupled Yield.

Objective function	Number of deletions	Suggested deletions[1]	Objective function value[2]	%Maximum Growth
Glycerol yield	6	FBA1, TDH1, GDH3, YDR111C, PRO2, MTD1	0.51	1.2
	3	FBA1, TDH1, RIP1	0.49	13
Glycerol Biomass Product Coupled Yield	6	FBA1, PDA1, OSM1, PDC6, GCV1, GAP1	41.12	46.24
	3	FBA1, TDH1, PDA1	27.48	20.26

[1] Only few of the suggested strategies, with high objective function values are shown. OptGene found many strategies with different, but high objective function values. This tendency can be controlled by varying GA parameters.
[2] Units are: Yield in gram (gram glucose)$^{-1}$,
Biomass Product Coupled Yield in milli-gram (gram-glucose.hour)$^{-1}$

Table S-8.3. List of metabolites that were allowed to be secreted out during OptGene simulations.

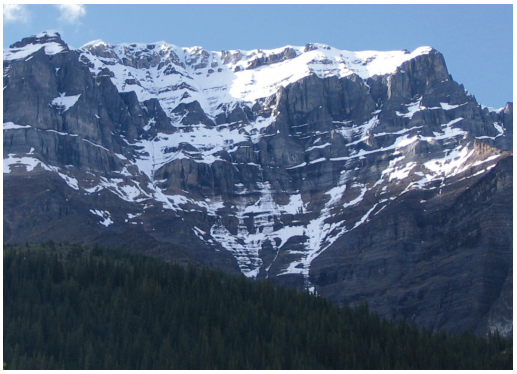
L-Glutamate, Glucose, 2-Oxoglutarate, Glycerol, L-Alanine, L-Arginine, L-Asparagine, L-Aspartate, L-Cysteine, Glycine, L-Glutamine, L-Histidine, L-Isoleucine, L-Leucine, L-Lysine, L-Methionine, L-Ornithine, L-Phenylalanine, L-Proline, L-Serine, L-Threonine, L-Tryptophan, L-Tyrosine, L-Valine, Guanine, HYXN, Xanthine, Acetate, Formate, Ethanol, Succinate, Urea, Ortho-phosphate, Citrate, Fumarate, (R)-Pantothenate, CO2, Acetaldehyde, Adenosine 3',5'-bisphosphate, dTTP, Thymine, D-Glucosamine 6-phosphate, 8-Amino-7-oxononanoate, Malate

Table S-8.4. Computational performance of OptGene as compared with the exhaustive search. Since CPU time will be dependent on the machine type and linear programming solver used, we have reported "search space" (or number of objective function evaluations) for Succinic acid case study. However, we note that the CPU time will be proportional to the number of objective function evaluations, independent of the machine type and solver. The fraction of independent runs converging to the best solution was taken into account while estimating the number of objective function evaluations for OptGene.

Number of deletions	Total search space	OptGene search space (Yield)	OptGene search space (BPCY)
3	2275280	115485	184677
4	1.35E+08	42712	452000
5	6.36E+09	103875	51750

Chapter 9: In silico metabolic engineering: experimental verification of predictions

The work presented in this chapter was done in collaboration with Donatella Cimini and Gaëlle Lettier.



"The rule is, jam tomorrow and jam yesterday - but never jam today"

The succinic acid case study discussed in the previous chapter was investigated experimentally for validation of the approach. Several single, double and triple deletion mutants were constructed and some of them were characterized for succinic acid production in batch and chemostat cultivations. Strain construction was started with the deletion of *SDH3* which was the common element between all of the deletion strategies generated by OptGene. We also performed detailed strain characterization including transcription analysis for this mutant (data not shown). Data obtained from chemostat cultures of the *sdh3Δ* mutant was used in OptGene in the form of additional constraints. Thus, the new multiple gene deletion predictions were generated with OptGene, by using both MOMA and FBA. The most promising strategy (additional deletions of *SER3-SER33* and *THR1*) was taken further. Construction of quadruple deletion mutant (*sdh3Δser3Δser33Δthr1Δ*) could not be achieved, possibly due to in-viability of the mutant. The *sdh3Δ* mutant and the double-reaction deletion mutants *(sdh3Δser3Δser33Δ* and *sdh3Δthr1Δ*) showed close to the predicted amount of succinate yields (data not shown). However, growth rates of the mutants were lower than predicted (data not shown). Moreover, *ser*3*Δ* and *ser*33*Δ* mutants were auxotrophic for serine. It was later possible to eliminate this auxotrophy through adaptive evolution experiments in shake flask cultivations (Otero JM, unpublished results). Based on the results obtained during this work, an integrative metabolic engineering strategy is proposed (Figure 9.1). The strategy involves repeated applications of experimental and computational techniques. With the aid of transcriptome and Metabolome data, such a strategy will help in achieving metabolic engineering targets in a rational and rapid way.

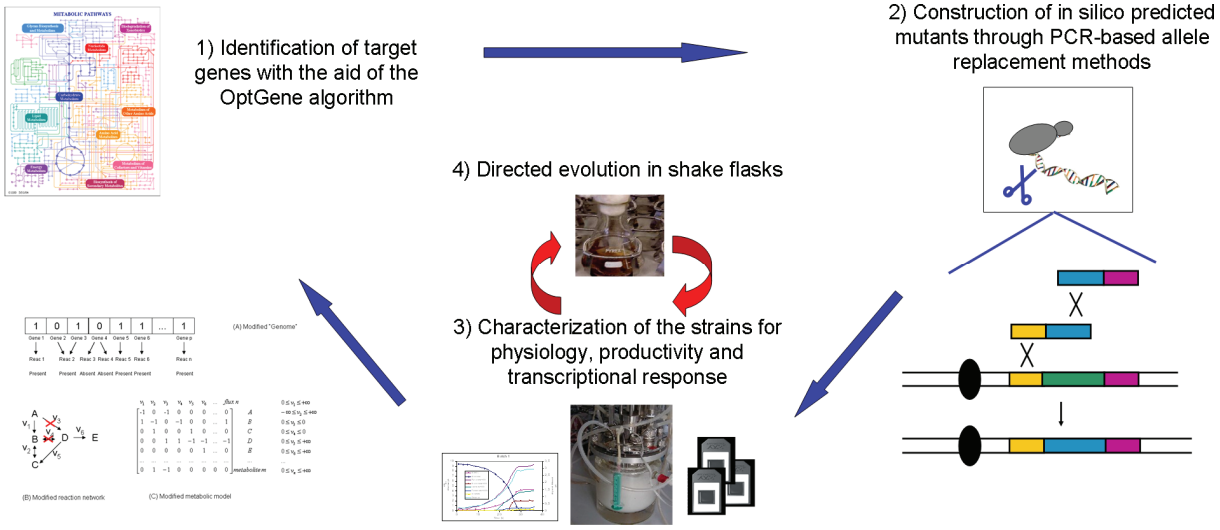


Figure 9.1. Proposed integrative metabolic engineering cycle for improved production of microbial metabolites. OptGene predictions, strain construction, characterization and adaptive evolution should be used in sequence. Several rounds of this cycle may be necessary for achieving high yield/productivity. Data generated during characterization (fermentation profiles, yields, transcriptome etc.) should be used together with OptGene in order to improve the prediction-confidence in the next round.

Chapter 10: Additional miscellaneous research

"Everything's got a moral, if only you can find it."

10.1 Reporter EFMs

Elementary flux modes (EFM) are stoichiometrically balanced sets of reactions that can operate at steady state without net production or consumption of intra-cellular metabolites (Schuster et al., 2000). EFM can also be seen as the stoichiometric definition of a pathway. Cellular metabolism can thus be decomposed into several EFMs. Weighted linear combinations of EFMs can be used to represent any steady network state. The usefulness of EFM-based network decomposition has not only stoichiometric but also regulatory implications (Stelling et al., 2002;Cakir et al., 2004). Moreover, EFMs also provide a conceptually easy way to understand the complex structure of metabolic network in the flux space.

Together with Intawat Nokaew (King Mongkut's University of Technology Thonburi, Thailand), I have tried to investigate whether transcriptional changes in metabolism can be explained by using EFM decomposition. The rational behind this strategy was that if fluxes through certain EFMs are more affected by a perturbation, those changes might be reflected at the transcriptional level. Since any EFM is stoichiometrically balanced, it is possible that some constraints are also placed on the expression of the corresponding genes. Thus, reporter EFMs are EFMs that show significant collective transcriptional response of all corresponding genes. The scoring procedure is the same as that for reporter metabolites. Indeed, we identified physiologically relevant EFMs following the deletion of *SDH3* and the shift from aerobic to anaerobic conditions. Figure 10.1 shows the distribution of reporter EFM scores for these two cases. EFMs were calculated for a small yeast metabolic model mainly comprising of the central carbon metabolism (around 80 reactions). It appears that the gene deletion lead to stronger expression changes as compared to the environmental change. It should be noted that deletion of *SDH3* also led to respiration deficient growth and hence the two perturbations are alike in some respect. However, detailed study of the calculated reporter EFMs is necessary before drawing any major conclusions. Also, it is necessary to consider that the two transcription studies were made at different growth conditions, *viz.*, one in chemostat and other in batch.

The concept of reporter EFM may be a very useful tool for integrating transcriptional data into flux estimations and analysis. However, since one reaction is typically participating in a large number of EFMs, reporter metabolites (Patil and Nielsen, 2005) (Chapter 3) and control effective fluxes (Stelling et al., 2002) may be more easy-to-interpret and use in general. Reporter EFM is an extension of the concept of reporter metabolites to pathway-scale. However, it is more difficult to imagine and hypothesize the evolutionary origin of such regulatory architecture. Hence al-

though the concept of Reporter EFM will be useful for data integration, it is unlikely that it will help us to discover new regulatory circuits.

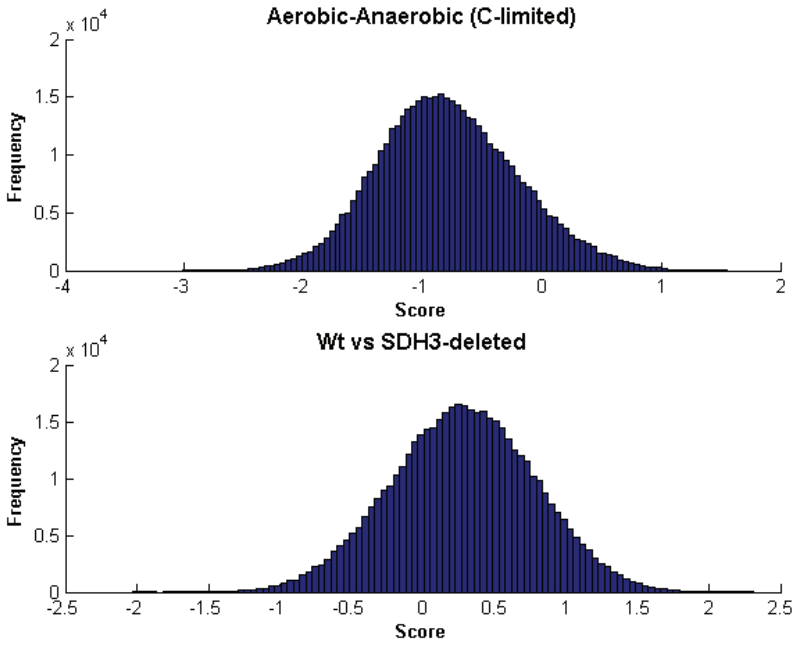


Figure 10.1 Distribution of transcriptional response score for EFMs following two different perturbations.

10.2 Reporter conditions and k-Cross algorithm

Regulatory response of cells is often characterized by time dependent changes in the expression of regulatory genes and consequently in their target genes. The systematic organization of transcriptional response around certain metabolites (Chapters 3, 6 and 7) may partially be attributed to transcription factors (TFs) binding to promoter regions of neighboring genes of a metabolite (Chapter 6). For some of the metabolites we have previously identified conserved promoter motifs for their neighboring genes and the TFs that are likely to bind at these motifs. However, for the rest of the metabolites there still may be some TFs/regulators that are directly/indirectly responsible for governing expression of metabolite-related genes. In this study such metabolite-related regulators have been identified through a correlation analysis of a time-series transcriptome data. The basic idea is to calculate the average correlation between expressions of all known/potential TFs/regulators and all neighbors of a metabolite. This correlation is then normalized by using correlations between TFs/regulators and randomly selected metabolic genes. This procedure would identify Tfs/regulators that are significantly correlated with the expression of the metabolite's neighbors. Thus, new regulatory proteins responsible for the coordinated transcriptional changes around a metabolite can be identified by the algorithm. Although the al-

gorithm is simple, there are some difficulties. Transcriptional changes in a regulatory cascade are, in general, highly non-linear. Therefore it is difficult to identify significant regulatory relationships by testing only for linear correlations. This difficulty can be partially overcome by: i) focusing only on reporter metabolites for a particular dataset (reporter metabolites imply good linear correlation around them, and hence it can be expected that the corresponding regulators, if any, are also linearly correlated with these genes); ii) Selecting the subset of experiments (/conditions) that maximizes the correlation score among the neighbor genes. The second heuristic is a challenging computational problem. Since correlation score (reporter score) is not only the function of expression of genes in question, but also the background distribution of scores, the problem of maximization of the score by selecting a subset of conditions is combinatorial complex. Genetic algorithm was used to address this problem (Figure 10.2). Although genetic algorithms do not guarantee to find the optimal solution, it usually results in a close to optimal solution which in this case may be sufficient. The conditions thus identified are defined as reporter conditions. Each reporter metabolite will be thus associated with reporter conditions. Since the background score calculation has to be repeated for each round of the genetic algorithm, the computational time required for the reporter condition algorithm was high (data not shown).

Once the reporter conditions are identified, the next step for finding the most relevant regulators/TFs is relatively easy. This algorithm is termed k-Cross algorithm (figure 10.3).

All potential TFs/regulators (including genes with unknown functions) are tested for correlation with the neighbors of the reporter metabolites over the corresponding reporter conditions. The top k regulators for each metabolite are then selected as putative regulators for the corresponding genes. The parameter k can either be decided arbitrarily (e.g. 10, 20 etc.) or based on a statistical significance cut-off (e.g. p-value of 0.05). An example of an outcome of the k-Cross algorithm is shown in figure 10.4, where the metabolite nodes (red) are connected to the identified putative regulatory proteins (blue nodes). The data used for this example is from the metabolic cycle experiment (Tu et al., 2005). Yeast cells were grown in a chemostat and synchronized oscillations were introduced with a pulse of glucose. Transcription of all genes was then measured over three cycles. The reasons for choosing this particular dataset are: i) metabolic cycle is closely related with the cell cycle; ii) oscillations during the metabolic cycle are natural in the sense that they are not introduced or synchronized through the addition of external agents; iii) dataset includes large number of time-point measurements (36). Interestingly, the top scoring reporter metabolite from this dataset is NADH, which is a well known key player in yeast metabolism which is also related to metabolic oscillations.

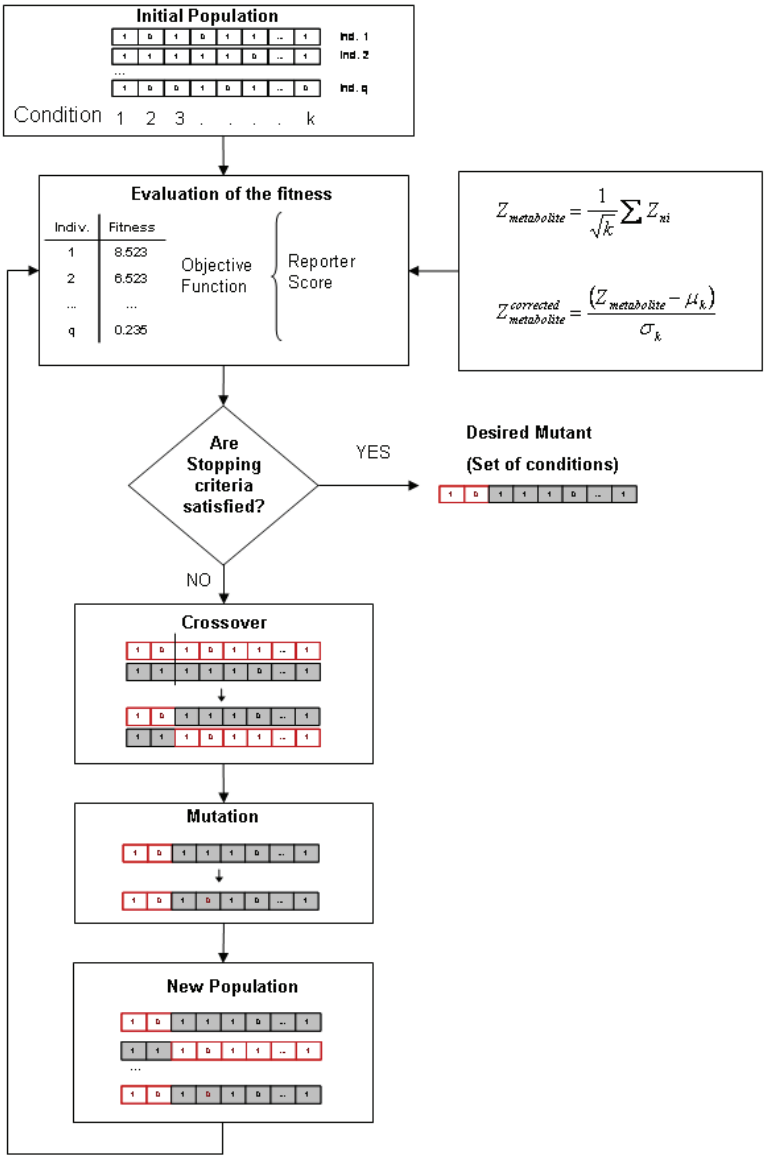


Figure 10.2. Genetic algorithm used for identifying reporter conditions for reporter metabolites.

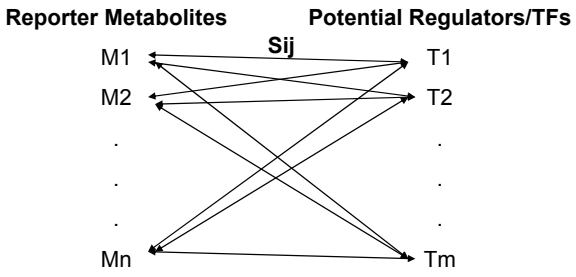


Figure 10.3. Schematic overview of the k-Cross algorithm.

The high connectivity in the network depicted in figure 10.4 implies that the regulatory mechanisms underlying changes in the metabolism are closely interrelated. Only two reporter metabolites appear to be outliers in this network. However, when k was increased to 20, the network became connected (data not shown). Indeed, regulators that can exert simultaneous control at several points will be an ideal control system for regulating highly connected networks such as metabolic network. Several of the identified potential regulators are genes with unknown functions or orphan ORFs. Thus the k-Cross algorithm can help in functional genomics by assigning functions to these genes. Experimental verification can be done for the high confidence interactions that emerge through the k-Cross algorithm applied to many different datasets.

In summary, the k-cross algorithm will help in reconstructing metabolic regulatory circuits and their links to other cellular processes through a hypothesis-driven modeling approach. Moreover, the k-cross algorithm will help in expanding the genome-scale metabolic models and their applicability over larger number of genes and hence to a larger number of environmental and genetic conditions.

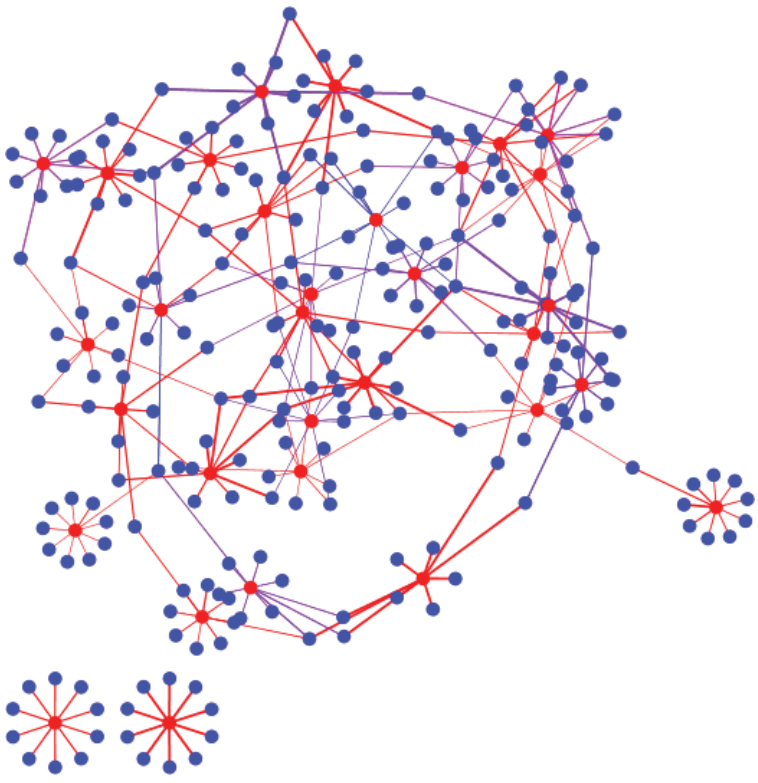


Figure 10.4. Example of an output of the reporter conditions and the k-Cross algorithm with k=10. The red nodes represent metabolites while blue nodes represent TFs/Regulators. The thickness of an edge is proportional to the correlation strength. The color of an edge is dependent on the number of positive correlations between the regulator and the neighbor genes of the metabolite. Edge with the red color indicates that all corresponding cor-

relations are positive, while an edge with the blue color implies all negative correlations. For the edges representing mixed positive and negative correlations, the color of an edge is a mixture of red and blue.

10.3 Essentiality of genes around metabolites

Several of the yeast genes are essential for growth, meaning that the deletion of any of these genes is lethal. Evolutionary and functional explanations for the question, why certain genes are essential while other are dispensable, has been a recent focus of research (Papp et al., 2004;Samal et al., 2006). The usefulness of a metabolite-centered analysis for studying stoichiometric and regulatory aspects of the yeast metabolic network has been demonstrated in the previous chapters. On the same lines, I queried whether the essential genes are preferentially clustered around certain metabolites. The results of this analysis are shown in table 10.1. As hypothesized by (Samal et al., 2006) many low degree metabolites were found to harbor essential genes. Notably, several highly connected metabolites were also found to be significant. E.g., the vital role of ATP in a variety of cellular functions is reflected by the significant concentration of essential genes around it. On the other hand H^{+}_{Ext} (extra-cellular proton, involved in several transport reactions) and glucose were found to have significantly lower fraction of essential genes around them. This is the consequence of the fact that several transport functionalities are backed up by isozymes and thus deletion of a single gene does not lead to an inviable phenotype. These results increase the dimensions of biological information that can be unraveled through the reporter approach and further support the hypothesis of the metabolite-centered evolution of metabolic networks.

10.4 Genome positioning of metabolite's neighbor genes

Potential regulatory and evolutionary mechanisms behind the observed transcriptional coregulation between metabolically related genes have been discussed in chapter 6. In this study, another dimension is added to these efforts by identifying the metabolites whose neighbor genes are significantly closely positioned in the genome. The hypothesis behind this investigation is: if the neighbor genes of a certain metabolite are closely positioned in a genome, it may partially explain the coregulation between, and/or the evolutionary origin for, these genes. For example, the neighbor genes of the metabolite under question might have been acquired from other species via a horizontal DNA transfer; or some of these genes have their origin in a gene duplication event, or the coregulation among these genes might be linked to their close positioning in the genome. Table 10.2 lists such metabolites whose neighboring genes are significantly closely located in the genome. Interestingly, ATP and redox co-factors make their way in this list. This fact further highlights the functional importance of redox and energy co-factors for metabolic networks as discussed in chapter 7.

Table 10.1. Reporter metabolites for gene essentiality. Yellow shaded metabolites indicate that the number of essential reactions around them is less than expected.

Metabolite	Number of neighbors	Essential neighbors	p-value
H+EXT	188	2	7.83367E-08
Pyrophosphate	68	20	1.61167E-06
ATP	164	34	1.74369E-06
(R)-5-Phosphomevalonate	5	5	8.83289E-06
Isopentenyl diphosphate	4	4	9.24372E-05
1-Phosphatidyl-D-myo-inositol	8	5	0.000367057
3-Dehydrosphinganine	3	3	0.00095731
2-Amino-4-hydroxy-6-hydroxymethyl 7,8-dihydropteridine	3	3	0.00095731
(R)-Mevalonate	7	4	0.002392806
AMP	46	11	0.002614205
ADP	129	22	0.002677518
1-Phosphatidyl-1D-myo-inositol 4-phosphate	4	3	0.00345949
alpha-D-Mannose 1-phosphate	2	2	0.009812425
Palmitoyl-CoA	2	2	0.009812425
(R)-5-Diphosphomevalonate	2	2	0.009812425
Dimethylallyl diphosphate	2	2	0.009812425
Geranyl diphosphate	2	2	0.009812425
trans,trans-Farnesyl diphosphate	2	2	0.009812425
Squalene	2	2	0.009812425
(S)-2,3-Epoxysqualene	2	2	0.009812425
Intermediate_Methylzymosterol_I	2	2	0.009812425
Intermediate_Zymosterol_I	2	2	0.009812425
N-Acetyl-D-glucosamine 1-phosphate	2	2	0.009812425
(R)-3-Hydroxy-3-methyl-2-oxobutanoateM	2	2	0.009812425
(R)-2,3-dihydroxy-3-methylbutanoateM	2	2	0.009812425
L-1-Pyrroline-3-hydroxy-5-carboxylate	2	2	0.009812425
trans-4-Hydroxy-L-proline	2	2	0.009812425
2-Amino-7,8-dihydro-4-hydroxy-6-(diphosphooxymethyl)pteridine	2	2	0.009812425
Porphobilinogen	2	2	0.009812425
Hydroxymethylbilane	2	2	0.009812425
Coproporphyrinogen	2	2	0.009812425
dUMP	6	3	0.014113231
alpha-D-Glucose	38	0	0.017016055
NADH	52	1	0.022184111
beta-D-Glucose 6-phosphate	3	2	0.026565345
CDPdiacylglycerol	3	2	0.026565345
Lanosterol	3	2	0.026565345
Thiamin diphosphate	3	2	0.026565345
4-Aminobenzoate	3	2	0.026565345
Dihydropteroate	3	2	0.026565345
Uroporphyrinogen III	3	2	0.026565345
NAD+	58	2	0.0426232
2-Oxoglutarate	29	0	0.045411636
CDPdiacylglycerolM	4	2	0.047941455
dTMP	4	2	0.047941455
Dihydrofolate	4	2	0.047941455
FMN	4	2	0.047941455

Table 10.2. Reporter metabolites for genome position of neighboring genes.

Metabolite	Position P-value
ATP	7.65304E-05
L-Glutamate	0.000659393
D-Fructose	0.001454756
NAD+M	0.001630868
alpha-D-Glucose	0.004812615
trans,trans-Farnesyl diphosphate	0.005645037
NAD+	0.006066368
FADM	0.006902072
3-Methyl-2-oxobutanoateM	0.007672133
1L-myo-Inositol 1-phosphate	0.007850773
FADH2M	0.008488164
D-Glyceraldehyde 3-phosphate	0.009626906
Orthophosphate	0.010951956
NADHM	0.012071059
2-Oxoglutarate	0.013339589
Oxygen	0.01398382
D-Ribulose 5-phosphate	0.013997373
3-Oxoacyl-CoA	0.014152249
D-Fructose 2,6-bisphosphate	0.01456584
dCTP	0.014742149
Phosphatidyl-N-methylethanolamine	0.017849471
Allantoate	0.019567243
beta-D-Fructose 6-phosphate	0.024453588
Malonyl-[acyl-carrier protein]	0.026619947
Acyl-[acyl-carrier protein]	0.026619947
CO2M	0.02728967
L-Cysteine	0.030544551
L-Cystathionine	0.031036076
AMP	0.034060165
H2O2	0.038737366
Hydrogen sulfide	0.039111586
L-Lysine	0.040969911
L-Homoserine	0.041440415
Choline phosphate	0.043078276
dUTP	0.044844549
NADH	0.048648642
Fumarate	0.04981706

10.5 Nonlinear correlation test for the analysis of the transcriptomics data

In this study, it is hypothesized that the expression profiles of the genes with regulatory relationship can be modeled using the Hill equation. This hypothesis is used to analyze dynamics of gene expression during beer fermentation. Since it is likely that many transcriptional regulatory circuits involve non-linear correlations, the here-reported method can be used to extract such interactions from the genome-scale transcriptome data.

Genome-wide gene expression datasets provide an opportunity to uncover complex transcriptional regulatory networks. One of the challenges in computational biology is to effectively extract regulatory information from these datasets. Presently, the most commonly used methods involve statistical significance tests and clustering algorithms. Most of the clustering algorithms use Pearson correlation coefficient (or some other measure of linear dependency tests) as a measure of the distance between gene expression profiles. In other words, these methods hypothesize a linear correlation between expressions of two genes. Although this hypothesis may be true for some genes, large number of genes might also have a non-linear correlation between their expression patterns. This information might be very valuable for correctly deducing regulatory network structures from the transcriptome data.

Here a method is reported to test the sigmoidal correlation between the gene expression profiles. The reason for choosing the sigmoidal relationship is that many enzyme regulatory systems have been shown to exhibit a sigmoidal response which characterizes a robust regulatory system (Mutalik et al., 2003;Koshland, Jr., 1998). The sigmoidal response was modeled by using the Hill equation,

$$f = \frac{I^n}{K + I^n}$$

Where n is termed as Hill coefficient. For n = 1, the system shows Michaelis-Menten behavior, for n>1 system shows ultrasensitive response and for n<1 the response is subsensitive (Fig. 10.5). The nonlinear regression between two gene expression profiles was performed using MATLAB.

As a case study, gene expression data from *Saccharomyces carlsbergensis*, obtained during production-scale lager beer fermentation was analyzed. The data included 12 samples taken in the time course of batch fermentation at the intervals of 1 day (with the exception of the second sample, which was taken 1 hour after the start of the fermentation). Due to computation limitations it was not possible to compare the expression profile between all genes. Therefore only the expressions of the genes that are known to be involved in glucose repression were selected for the analysis.

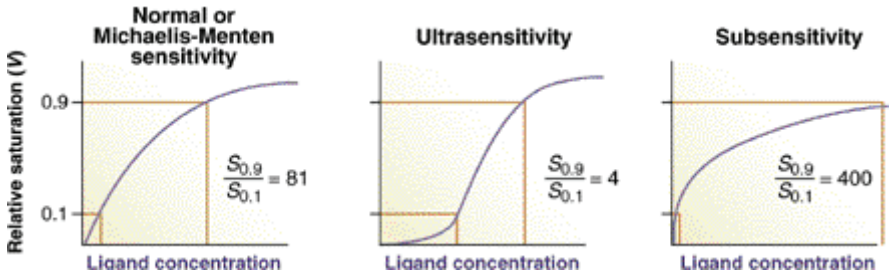


Figure 10.5. Normal, subsensitive and ultrasensitive responses.

Thus, non-linear regression coefficients were calculated for the genes known to be involved in glucose repression pathway, against rest of the genes. The top scoring hundred correlations were chosen for further analysis.

Some of the interesting correlations include:

1) Biosynthetic genes involved in the fermentative metabolism (alcohol dehydrogenase *ADH4*, DL-glycerol phosphate phosphatase *HOR2*, 3-isopropylmalate dehydrogenase *LEU2*).

2) Genes involved in the energy metabolism (ATPase in the plasma membrane *PAM1*, mitochondrial ATP-synthase subunit *ATP2*, vacuolar ATPase *VMA5*).

3) Genes involved in the cell cycle (*SIM1*).

4) Other genes involved in the glucose repression (*SNF4*).

A few examples of the identified correlations are shown in Figure 10.6.

The described results could not be obtained by using classical methods that are based on linear correlation coefficient. The expression profiles of several genes, which were found to show good sigmoidal correlation with *RGT2* are displayed in Figure 10.7. A simple visual inspection would tell that the genes do not have a linear correlation and hence cannot be grouped by using classical clustering methods.

In conclusion, here-described approach can be used to systematically identify non-linearly correlated bio-molecular interactions from omics data. This will expand the scope of many existing algorithms, including the reporter algorithm, for identifying and taking advantage of previously unknown interactions.

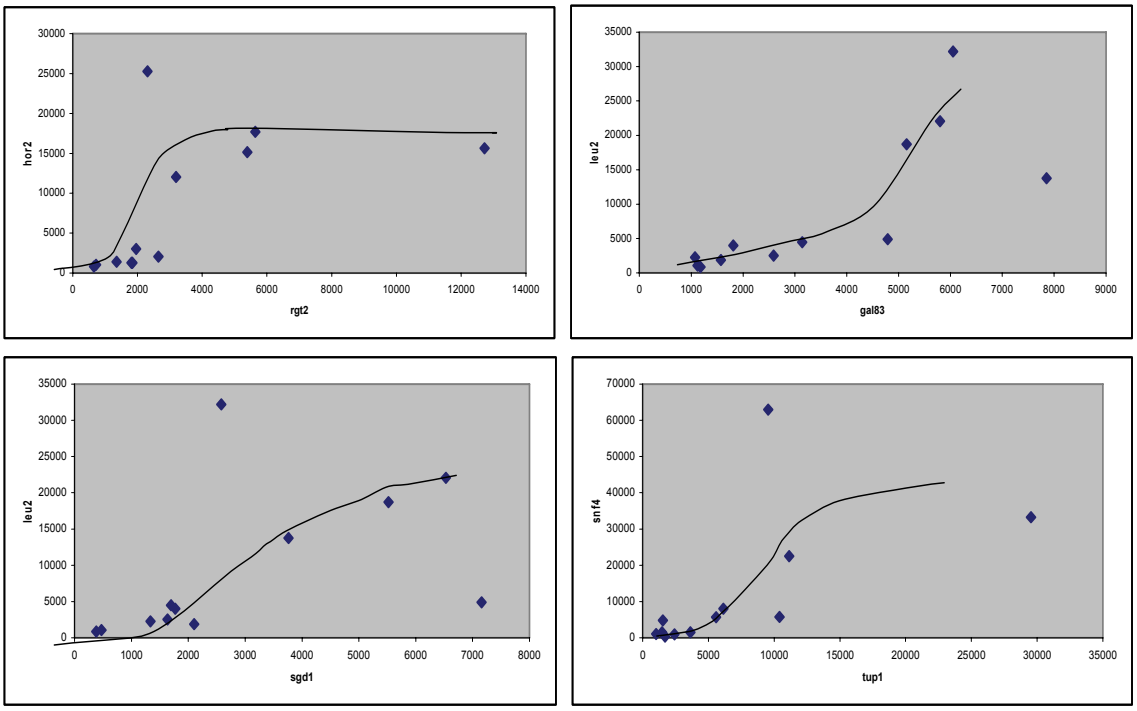


Figure 10.6. Expression profiles for some of the identified gene-pairs with sigmoidal correlation.

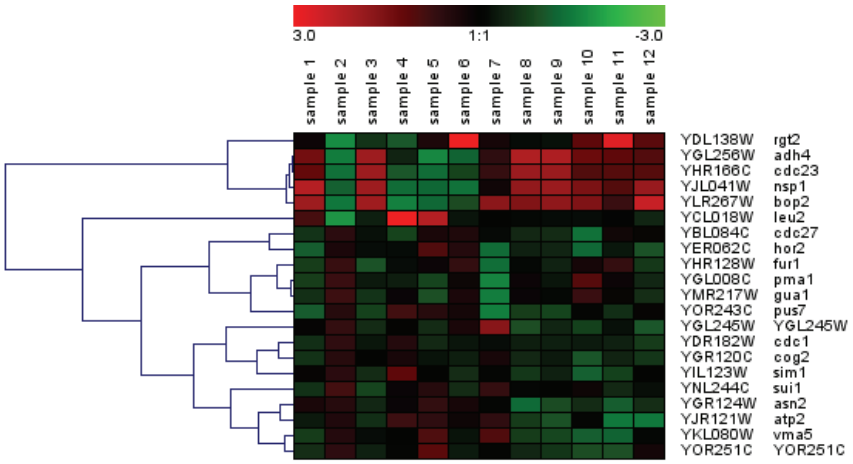


Figure 10.7. Expression profiles of the genes that were found to have sigmoidal correlation with the expression of *RGT2*.

Chapter 11: Conclusions and future perspectives

"Would you tell me please, which way I ought to go from here?" "That depends a good deal on where you want to get to," said the Cat. "I don't care much where--" said Alice. "Then it doesn't matter which way you go," said the Cat.

Evolutionary insight into emergence and organization of metabolic regulation is one of the most promising leads from this work. It appears that the regulation in a metabolic network is guided by the topology of the network itself. Neighboring genes of a metabolite (meaning the genes coding for the enzymes that catalyze reactions involving that metabolite) were thus found to be transcriptionally coregulated following perturbations in that part of the metabolism. The metabolites that harbor significant transcriptional coregulation in a particular experiment were termed as reporter metabolites. Reporter metabolites unravel a simple principle underlying complex regulatory response in a metabolic network. This principle enables us to understand the logic of the cellular response to genetic/environmental perturbations without knowing the architecture of regulatory machinery implementing these changes. It is argued that the transcriptional coregulation of genes surrounding a metabolite is partly a thermodynamic necessity for the cells, in order to either maintain the homeostasis or to adjust the metabolic fluxes and pools to new demands. Notably, reporter metabolites were found to be perturbation-specific and unbiased towards the connectivity. Accordingly, the importance of the highly connected metabolic co-factors in terms of integrating the operation of distinct metabolic functions was also found to be reflected in the coregulation of their neighbor genes. Furthermore, from a regulatory point of view, metabolite-based grouping of genes was found to be more significant as compared with the traditional pathway-based grouping.

The concept of reporter metabolites was further enriched by identifying specific mechanisms responsible for the organization of transcriptional coregulation around them. Several metabolically related gene groups were found to be evolutionary conserved and also transcriptionally co-regulated. Notably, these gene groups also display common sequence motifs in their promoter regions. These results imply that the regulatory circuits have evolved around conserved and metabolically related genes. Transcriptional evidences were found for some of the new potential regulatory circuits identified by the reporter analysis. Owing to the foundation of the reporter approach based on the network topology, which is conserved in many species, principles of regulation can now be easily and systematically generalized, categorized and extrapolated across species. Indeed, the reporter metabolite algorithm has already been successfully used to decipher regulatory information in several systems including humans (Figure 11.1). Together with the fact that the metabolic networks are widely conserved from bacteria to humans, reporter algorithm presents an opportunity to explore the regulatory principles of human metabolism. Thus, knowledge gained from microbial metabolism can be systematically transferred and enriched for understanding the basis of metabolic diseases in humans. An overview of an example (proposed) strategy that can be applied for metabolic disease research is shown in figure 11.2.

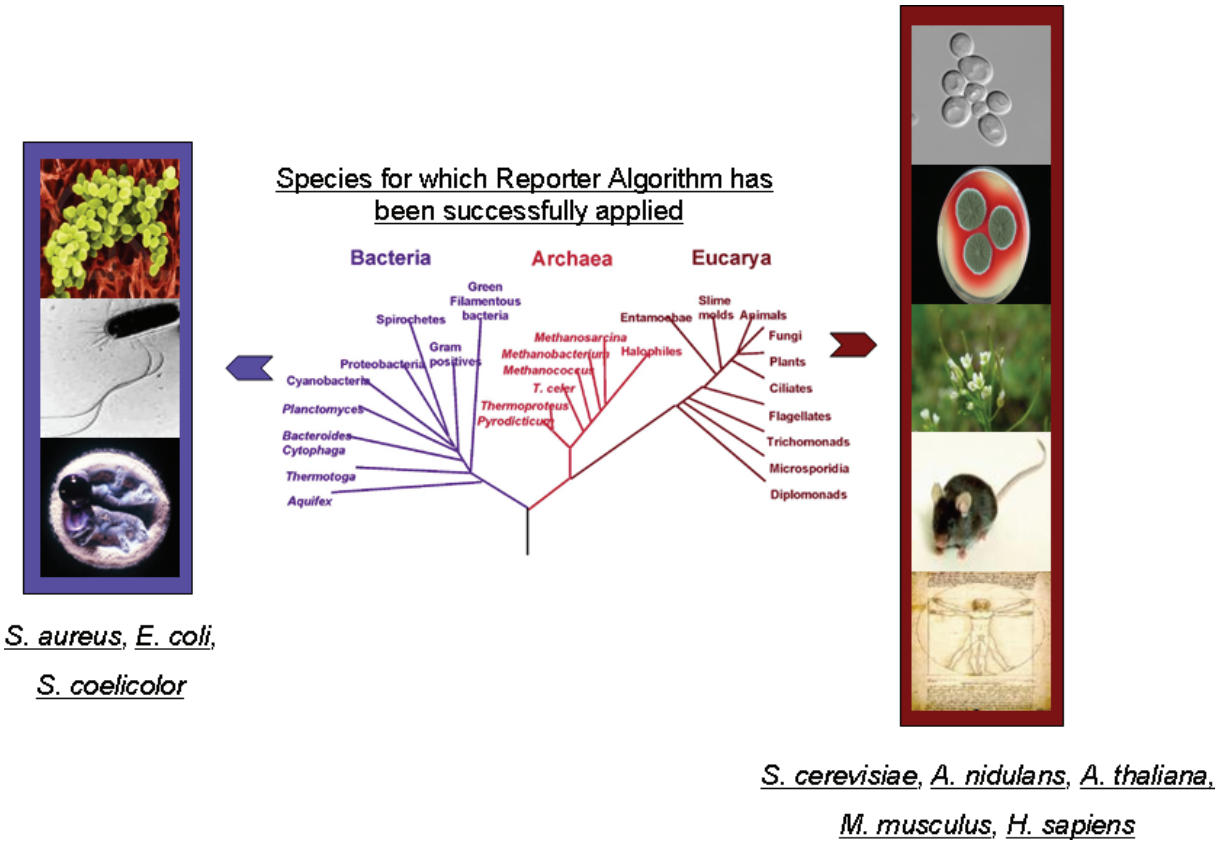


Figure 11.1. Schematic overview of different species for which Reporter Metabolite and Metabolic Sub-network Algorithm has been successfully applied for uncovering the principles of transcriptional regulation. The list includes several industrially important cell-factories such as *Streptomyces coelicolor*, *Aspergillus niger*, *Saccharomyces cerevisiae* and *Escherichia coli*. Remarkably, transcriptome analysis for higher eukaryotes including plants, mouse and human has also shown good promise. Consequently, further development of the algorithm will help in understanding the metabolic basis of human diseases such as diabetes and cancer; and in medical biotechnology for understanding pathogenesis. The wide applicability of this novel algorithm across several domains of life, ranging from bacteria to humans, illustrates the power of system level analysis and helps to connect the research and researchers in different fields of biology.

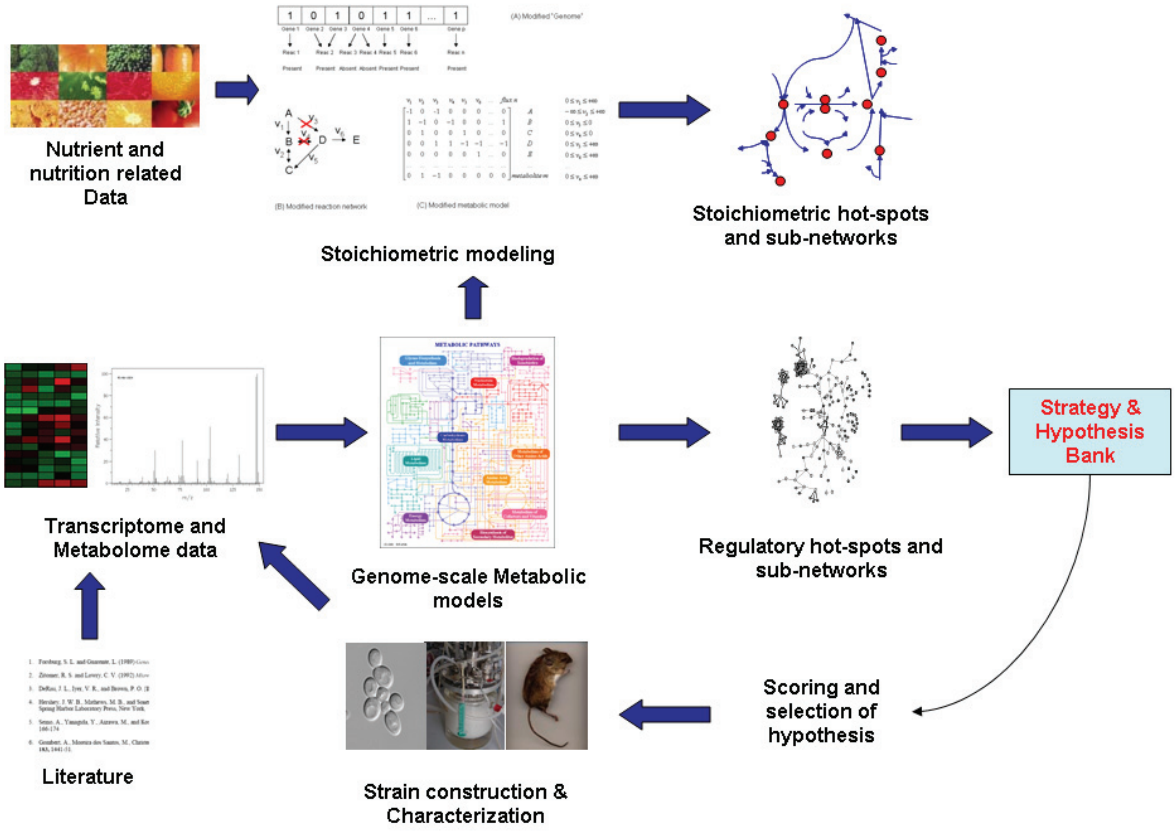


Figure 11.2. Schematic representation of the proposed research strategy for metabolic disease research. Omics data from human, mouse and yeast together with the state of the art computational tools can be used to generate concrete testable hypotheses for uncovering the basis of metabolic disorders. Selected hypotheses then can be used to design new strains and experiments. Results from those may then be used to iteratively improve the model and consequently new hypotheses and conclusions will be generated.

Problem of the evolution of metabolic networks and associated regulatory networks needs further detailed investigation. Although a simple explanation of metabolite centered emergence of regulation can explain several observations, far more complicated examples of regulatory circuits exist even in bacteria. How simple regulatory rules are combined together to create complex circuits and what the rational behind them is will be a pressing question in the near future.

Extrapolation of the principles learned from metabolic networks to other bio-molecular interaction networks (such as protein-protein and protein-DNA interaction networks) will also be a natural follow up to this work. Indeed, generalization of the reporter metabolite concept to "reporter features" has been found very useful (AP Oliveira *et al.*, manuscript submitted). We have hypothesized that the topology of biological interactions (either physical or functional) itself guides (and constrains) the regulatory response of the network following a perturbation of the

system. The simplest form of a regulatory principle stemming from this hypothesis is that the regulatory response starts at the first neighbors of a node where the perturbation is introduced, or which is most affected by the perturbation (figure 11.3). This response can then subsequently spread to the next neighbors and so on. This hypothesis can be used to understand the logic behind the action of the cellular regulatory mechanisms by identifying key regulatory nodes around which the response is significantly concentrated.

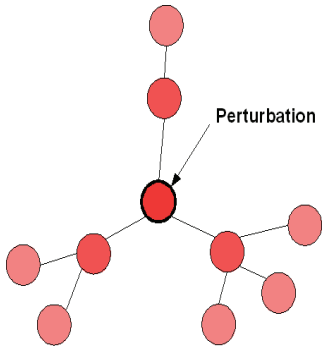


Figure 11.3. Hypothesis about how regulatory response is organized in biological interaction networks. Response first initiates in the neighboring nodes of the perturbed node. This response then may propagate to the second neighbors and so on.

Once the regulatory rules and architecture surrounding metabolism and other cellular processes are identified, the next step would be to integrate this knowledge with stoichiometric approaches in order to improve the predictive power of cellular modeling. Such an integrative approach will be useful not only for designing microbial cells *in silico* and then testing *in vivo* (synthetic biology), but also for finding cures for human diseases and predicting effects of drugs on cellular metabolism. This will also require a comprehensive modeling platform that combines together both stoichiometric and regulatory aspects of metabolic networks. OptGene algorithm (Chapter 8) will be a suitable starting point where regulatory modules are gradually and systematically added. Thus, transcriptome, metabolome and fluxome will find there way into algorithms for the identification of metabolic engineering targets and metabolic disease remedies.

I envision that the simple rules of regulation and operation of cellular processes can be enlisted in a biological "objects" database, much like objects used in the computer programming. Such objects can then be custom-modified and fitted together to create a phenotype of choice or incorporated into existing systems to repair defects. I see this thesis as a step towards this goal.

References

1. Agarwal, A. K. and R. J. Auchus. 2005. Minireview: Cellular Redox State Regulates Hydroxysteroid Dehydrogenase Activity and Intracellular Hormone Potency. *Endocrinology* 146:2531-2538.

2. Akesson, M., J. Forster, and J. Nielsen. 2004. Integration of gene expression data into genome-scale metabolic models. *Metab Eng* 6:285-293.

3. Albert, R., H. Jeong, and A. L. Barabasi. 2000. Error and attack tolerance of complex networks. *Nature* 406:378-382.

4. Allen, J., H. M. Davey, D. Broadhurst, J. K. Heald, J. J. Rowland, S. G. Oliver, and D. B. Kell. 2003. High-throughput classification of yeast mutants for functional genomics using metabolic footprinting. *Nat Biotechnol.* 21:692-696.

5. Anderson, R. M., M. Latorre-Esteves, A. R. Neves, S. Lavu, O. Medvedik, C. Taylor, K. T. Howitz, H. Santos, and D. A. Sinclair. 2003. Yeast Life-Span Extension by Calorie Restriction Is Independent of NAD Fluctuation. *Science* 302:2124-2126.

6. Beard, D. A., S. D. Liang, and H. Qian. 2002. Energy balance for analysis of complex metabolic networks. *Biophys. J* 83:79-86.

7. Bernard, F. and B. Andre. 2001. Ubiquitin and the SCF(Grr1) ubiquitin ligase complex are involved in the signalling pathway activated by external amino acids in Saccharomyces cerevisiae. *FEBS Lett.* 496:81-85.

8. Bino, R. J., R. D. Hall, O. Fiehn, J. Kopka, K. Saito, J. Draper, B. J. Nikolau, P. Mendes, U. Roessner-Tunali, M. H. Beale, R. N. Trethewey, B. M. Lange, E. S. Wurtele, and L. W. Sumner. 2004. Potential of metabolomics as a functional genomics tool. *Trends Plant Sci.* 9:418-425.

9. Bork, P. 2005. Is there biological research beyond Systems Biology? A comparative analysis of terms. *Mol Syst Biol* 1:msb4100016-msb41000E1.

10. Borodina, I. and J. Nielsen. 2005. From genomes to in silico cells via metabolic networks. *Curr. Opin. Biotechnol.* 16:350-355.

11. Botstein, D., S. A. Chervitz, and J. M. Cherry. 1997. GENETICS: Yeast as a Model Organism. *Science* 277:1259-1260.

12. Boucher, A., D. Lu, S. C. Burgess, S. Telemaque-Potts, M. V. Jensen, H. Mulder, M. Y. Wang, R. H. Unger, A. D. Sherry, and C. B. Newgard. 2004. Biochemical Mechanism of Lipid-induced Impairment of Glucose-stimulated Insulin Secretion and Reversal with a Malate Analogue. *J. Biol. Chem.* 279:27263-27271.

13. Brekasis, D. and M. S. B. Paget. 2003. A novel sensor of NADH/NAD(+) redox poise in Streptomyces coelicolor A3(2). *Embo Journal* 22:4856-4865.

14. Brem, R. B., G. Yvert, R. Clinton, and L. Kruglyak. 2002. Genetic dissection of transcriptional regulation in budding yeast. *Science* 296:752-755.

15. Bro, C., B. Regenberg, and J. Nielsen. 2004. Genome-wide transcriptional response of a Saccharomyces cerevisiae strain with an altered redox metabolism. *Biotechnol. Bioeng.* 85:269-276.

16. Bro, C. 2003. Transcription Analysis of the Yeast *Saccharomyces cerevisiae*. Technical University of Denmark.

17. Burgard, A. P. and C. D. Maranas. 2001. Probing the performance limits of the Escherichia coli metabolic network subject to gene additions or deletions. *Biotechnol. Bioeng.* 74:364-375.

18. Burgard, A. P. and C. D. Maranas. 2003. Optimization-based framework for inferring and testing hypothesized metabolic objective functions. *Biotechnol. Bioeng.* 82:670-677.

19. Burgard, A. P., P. Pharkya, and C. D. Maranas. 2003. Optknock: A bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* 84:647-657.

20. Burgard, A. P., E. V. Nikolaev, C. H. Schilling, and C. D. Maranas. 2004. Flux Coupling Analysis of Genome-Scale Metabolic Network Reconstructions. *Genome Res.*1926504.

21. Cakir, T., B. Kirdar, and K. O. Ulgen. 2004. Metabolic pathway analysis of yeast strengthens the bridge between transcriptomics and metabolic networks. *Biotechnol. Bioeng.* 86:251-260.

22. Carlson, R., D. Fell, and F. Srienc. 2002. Metabolic pathway analysis of a recombinant yeast for rational strain development. *Biotechnol. Bioeng.* 79:121-134.

23. Chatterjee, R. and L. Yuan. 2006. Directed evolution of metabolic pathways. *Trends Biotechnol.* 24:28-38.

24. Christensen, B. and J. Nielsen. 2000. Metabolic network analysis. A powerful tool in metabolic engineering. *Adv. Biochem. Eng Biotechnol.* 66:209-231.

25. Coleman, S. T., T. K. Fang, S. A. Rovinsky, F. J. Turano, and W. S. Moye-Rowley. 2001. Expression of a glutamate decarboxylase homologue is required for normal oxidative stress tolerance in Saccharomyces cerevisiae. *J. Biol. Chem.* 276:244-250.

26. Covert, M. W., E. M. Knight, J. L. Reed, M. J. Herrgard, and B. O. Palsson. 2004. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 429:92-96.

27. Covert, M. W., C. H. Schilling, and B. O. Palsson. 2001. Regulation of Gene Expression in Flux Balance Models of Metabolism. *Journal of Theoretical Biology* 213:73-88.

28. Covert, M. W. and B. O. Palsson. 2002. Transcriptional Regulation in Constraints-based Metabolic Models of Escherichia coli. *J. Biol. Chem.* 277:28058-28064.

29. Covert, M. W. and B. O. Palsson. 2003. Constraints-based models: Regulation of Gene Expression Reduces the Steady-state Solution Space. *Journal of Theoretical Biology* 221:309-325.

30. Crick, F. 1970. Central Dogma of Molecular Biology. *Nature* 227:561-563.

31. Cronwright, G. R., J. M. Rohwer, and B. A. Prior. 2002. Metabolic Control Analysis of Glycerol Synthesis in Saccharomyces cerevisiae. *Appl. Environ. Microbiol.* 68:4448-4456.

32. Dandekar, T., F. Moldenhauer, S. Bulik, H. Bertram, and S. Schuster. 2003. A method for classifying metabolites in topological pathway analyses based on minimization of pathway number. *Biosystems* 70:255-270.

33. Daran-Lapujade, P., M. L. A. Jansen, J. M. Daran, W. van Gulik, J. H. de Winde, and J. T. Pronk. 2004. Role of Transcriptional Regulation in Controlling Fluxes in Central Carbon Metabolism of Saccharomyces cerevisiae: A CHEMOSTAT CULTURE STUDY. *J. Biol. Chem.* 279:9125-9138.

34. David, H., M. Akesson, and J. Nielsen. 2003. Reconstruction of the central carbon metabolism of Aspergillus niger. *Eur. J Biochem.* 270:4243-4253.

35. Dekel, E. and U. Alon. 2005. Optimality and evolutionary tuning of the expression level of a protein. *Nature* 436:588-592.

36. DeLuna, A., A. Avendano, L. Riego, and A. Gonzalez. 2001. NADP-Glutamate Dehydrogenase Isoenzymes of Saccharomyces cerevisiae. Purification, kinetic properties, and physiological roles. *J. Biol. Chem.* 276:43775-43783.

37. DeRisi, J. L., V. R. Iyer, and P. O. Brown. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278:680-686.

38. Devantier, R., B. Scheithauer, S. G. Villas-Boas, S. Pedersen, and L. Olsson. 2005a. Metabolite profiling for analysis of yeast stress response during very high gravity ethanol fermentations. *Biotechnol. Bioeng.* 90:703-714.

39. Devantier, R., S. Pedersen, and L. Olsson. 2005b. Transcription analysis of S. cerevisiae in VHG fermentation: The genome-wide transcriptional response of Saccharomyces cerevisiae during very high gravity ethanol fermentations is highly affected by the stationary phase. *Industrial Biotechnology* 1:51-63.

40. dos Santos, M. M., A. K. Gombert, B. Christensen, L. Olsson, and J. Nielsen. 2003. Identification of in vivo enzyme activities in the cometabolism of glucose and acetate by Saccharomyces cerevisiae by using 13C-labeled substrates. *Eukaryot. Cell* 2:599-608.

41. Edwards, J. S., R. U. Ibarra, and B. O. Palsson. 2001. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nat. Biotechnol.* 19:125-130.

42. Edwards, J. S. and B. O. Palsson. 2000. The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci. U. S. A* 97:5528-5533.

43. Edwards, J. S. and B. O. Palsson. 1999. Systems Properties of the Haemophilus influenzae Rd Metabolic Genotype. *J. Biol. Chem.* 274:17410-17416.

44. Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A* 95:14863-14868.

45. Erasmus, D. J., G. K. van der Merwe, and H. J. van Vuuren. 2003. Genome-wide expression analyses: Metabolic adaptation of Saccharomyces cerevisiae to high sugar stress. *FEMS Yeast Res.* 3:375-399.

46. Famili, I., J. Forster, J. Nielsen, and B. O. Palsson. 2003. Saccharomyces cerevisiae phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *PNAS* 100:13134-13139.

47. Fell, D. A. and J. R. Small. 1986. Fat synthesis in adipose tissue. An examination of stoichiometric constraints. *Biochem. J.* 238:781-786.

48. Fell, D. A. and A. Wagner. 2000. The small world of metabolism. *Nat. Biotechnol.* 18:1121-1122.

49. Ferea, T. L., D. Botstein, P. O. Brown, and R. F. Rosenzweig. 1999. Systematic changes in gene expression patterns following adaptive evolution in yeast. *Proc. Natl. Acad. Sci. U. S. A* 96:9721-9726.

50. Fernie, A. R., R. N. Trethewey, A. J. Krotzky, and L. Willmitzer. 2004. Metabolite profiling: from diagnostics to systems biology. *Nat Rev Mol. Cell Biol.* 5:763-769.

51. Flick, K. M., N. Spielewoy, T. I. Kalashnikova, M. Guaderrama, Q. Zhu, H. C. Chang, and C. Wittenberg. 2003. Grr1-dependent Inactivation of Mth1 Mediates Glucose-induced Dissociation of Rgt1 from HXT Gene Promoters. *Mol. Biol. Cell* 14:3230-3241.

52. Fong, S. S., J. Y. Marciniak, and B. O. Palsson. 2003. Description and Interpretation of Adaptive Evolution of Escherichia coli K-12 MG1655 by Using a Genome-Scale In Silico Metabolic Model. *J. Bacteriol.* 185:6400-6408.

53. Forster, J. 2003. Pathway analysis of the metabolic network of Saccharomyces cerevisiae. Technical University of Denmark.

54. Forster, J., I. Famili, P. Fu, B. O. Palsson, and J. Nielsen. 2003a. Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Res.* 13:244-253.

55. Forster, J., I. Famili, B. O. Palsson, and J. Nielsen. 2003b. Large-Scale Evaluation of *In Silico* Gene Deletions in *Saccharomyces cerevisiae*. *OMICS* 7:195-202.

56. Forster, J., A. K. Gombert, and J. Nielsen. 2002. A functional genomics approach using metabolomics and in silico pathway analysis. *Biotechnol. Bioeng.* 79:703-712.

57. Gancedo, J. M. and C. Gancedo. 1997. Gluconeogenesis and catabolite inactivation. *In* Yeast Sugar Metabolism: Biochemistry, Genetics, Biotechnology, and Applications. F. K. Zimeermann and K. D. Entian, editors. Technomic, Lancaster. 359-377.

58. Gertz, E. M. and S. J. Wright. 2003. Object-Oriented Software for Quadratic Programming. *ACM Transactions on Mathematical Software 29*58-81.

59. Goldberg, D. E. 1989. Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Reading, Massachussets.

60. Gombert, A. K. and J. Nielsen. 2000. Mathematical modelling of metabolism. *Current Opinion in Biotechnology* 11:180-186.

61. Goodacre, R., S. Vaidyanathan, W. B. Dunn, G. G. Harrigan, and D. B. Kell. 2004. Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends Biotechnol.* 22:245-252.

62. Hanisch, D., A. Zien, R. Zimmer, and T. Lengauer. 2002. Co-clustering of biological networks and gene expression data. *Bioinformatics* 18:145S-154.

63. Herrero, J., J. M. Vaquerizas, F. Al-Shahrour, L. Conde, A. Mateos, J. S. R. Diaz-Uriarte, and J. Dopazo. 2004. New challenges in gene expression data analysis and the extended GEPAS. *Nucl. Acids. Res.* 32:W485-W491.

64. Hirai, M. Y., M. Yano, D. B. Goodenowe, S. Kanaya, T. Kimura, M. Awazuhara, M. Arita, T. Fujiwara, and K. Saito. 2004. Integration of transcriptomics and metabolomics for understanding of global responses to nutritional stresses in Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U. S. A* 101:10205-10210.

65. Ibarra, R. U., J. S. Edwards, and B. O. Palsson. 2002. Escherichia coli K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. *Nature* 420:186-189.

66. Ideker, T. and D. Lauffenburger. 2003. Building with a scaffold: emerging strategies for high- to low-level cellular modeling. *Trends Biotechnol.* 21:255-262.

67. Ideker, T., V. Thorsson, J. A. Ranish, R. Christmas, J. Buhler, J. K. Eng, R. Bumgarner, D. R. Goodlett, R. Aebersold, and L. Hood. 2001. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292:929-934.

68. Ideker, T., O. Ozier, B. Schwikowski, and A. F. Siegel. 2002. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 18:233S-240.

69. Ihmels, J., G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, and N. Barkai. 2002. Revealing modular organization in the yeast transcriptional network. *Nat. Genet.* 31:370-377.

70. Ihmels, J., R. Levy, and N. Barkai. 2004. Principles of transcriptional control in the metabolic network of Saccharomyces cerevisiae. *Nat. Biotechnol.* 22:86-92.

71. Jansen, M. L., J. A. Diderich, M. Mashego, A. Hassane, J. H. de Winde, P. ran-Lapujade, and J. T. Pronk. 2005. Prolonged selection in aerobic, glucose-limited chemostat cultures of Saccharomyces cerevisiae causes a partial loss of glycolytic capacity. *Microbiology* 151:1657-1669.

72. Jeong, H., B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi. 2000. The large-scale organization of metabolic networks. *Nature* 407:651-654.

73. Jewett, M. C., G. Hofmann, and J. Nielsen. 2006. Fungal metabolite analysis in genomics and phenomics. *Current Opinion in Biotechnology* 17:191-197.

74. Kanehisa, M., S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. 2004. The KEGG resource for deciphering the genome. *Nucleic Acids Res.* 32 Database issue:D277-D280.

75. Kaniak, A., Z. Xue, D. Macool, J. H. Kim, and M. Johnston. 2004. Regulatory Network Connecting Two Glucose Signal Transduction Pathways in Saccharomyces cerevisiae. *Eukaryotic Cell* 3:221-231.

76. Kauffman, K. J., P. Prakash, and J. S. Edwards. 2003. Advances in flux balance analysis. *Curr. Opin. Biotechnol.* 14:491-496.

77. Kellis, M., B. W. Birren, and E. S. Lander. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. *Nature* 428:617-624.

78. Kharchenko, P., G. M. Church, and D. Vitkup. 2005. Expression dynamics of a cellular metabolic network. *Mol Syst Biol* 1:msb4100023-msb41000E1.

79. Klamt, S. and J. Stelling. 2002. Combinatorial complexity of pathway analysis in metabolic networks. *Mol. Biol. Rep.* 29:233-236.

80. Klamt, S. and J. Stelling. 2003. Two approaches for metabolic pathway analysis? *Trends Biotechnol.* 21:64-69.

81. Klapa, M. I., J. C. Aon, and G. Stephanopoulos. 2003. Systematic quantification of complex metabolic flux networks using stable isotopes and mass spectrometry. *Eur. J Biochem.* 270:3525-3542.

82. Koshland, D. E., Jr. 1998. BIOCHEMISTRY: Enhanced: The Era of Pathway Quantification. *Science* 280:852-853.

83. Kuffner, R., R. Zimmer, and T. Lengauer. 2000. Pathway analysis in metabolic databases via differential metabolic display (DMD). *Bioinformatics* 16:825-836.

84. Kummel, A., S. Panke, and M. Heinemann. 2006. Putative regulatory sites unraveled by network-embedded thermodynamic analysis of metabolome data. *Mol Syst Biol* 2.

85. Lee, I., S. V. Date, A. T. Adai, and E. M. Marcotte. 2004. A probabilistic functional network of yeast genes. *Science* 306:1555-1558.

86. Liao, J. C., S.-Y. Hou, and Y.-P. Chao. 1996. Pathway analysis, engineering, and physiological considerations for redirecting central metaboli**sm**. *Biotechnol. Bioeng.* 52:129-140.

87. Lin, S. J., P. A. Defossez, and L. Guarente. 2000. Requirement of NAD and SIR2 for Life-Span Extension by Calorie Restriction in Saccharomyces cerevisiae. *Science* 289:2126-2128.

88. Mahadevan, R., J. S. Edwards, and F. J. Doyle, III. 2002. Dynamic Flux Balance Analysis of Diauxic Growth in Escherichia coli. *Biophys. J.* 83:1331-1340.

89. McCammon, M. T., C. B. Epstein, B. Przybyla-Zawislak, L. Alister-Henn, and R. A. Butow. 2003. Global Transcription Analysis of Krebs Tricarboxylic Acid Cycle Mutants Reveals an Alternating Pattern of Gene Expression and Effects on Hypoxic and Oxidative Genes. *Mol. Biol. Cell* 14:958-972.

90. Mewes, H. W., C. Amid, R. Arnold, D. Frishman, U. Guldener, G. Mannhaupt, M. Munsterkotter, P. Pagel, N. Strack, V. Stumpflen, J. Warfsmann, and A. Ruepp. 2004. MIPS: analysis and annotation of proteins from whole genomes. *Nucl. Acids. Res.* 32:D41-D44.

91. Michalewicz, Z. 1996. Genetic Algorithms + Data Structures = Evolution Programs. Spinger.

92. Miki, R., K. Kadota, H. Bono, Y. Mizuno, Y. Tomaru, P. Carninci, M. Itoh, K. Shibata, J. Kawai, H. Konno, S. Watanabe, K. Sato, Y. Tokusumi, N. Kikuchi, Y. Ishii, Y. Hamaguchi, I. Nishizuka, H. Goto, H. Nitanda, S. Satomi, A. Yoshiki, M. Kusakabe, J. L. DeRisi, M. B. Eisen, V. R. Iyer, P. O. Brown, M. Muramatsu, H. Shimada, Y. Okazaki, and Y. Hayashizaki. 2001. Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays. *Proc. Natl. Acad. Sci. U. S. A* 98:2199-2204.

93. Moreira dos, S. M., G. Thygesen, P. Kotter, L. Olsson, and J. Nielsen. 2003. Aerobic physiology of redox-engineered Saccharomyces cerevisiae strains modified in the ammonium assimilation for increased NADPH availability. *FEMS Yeast Res.* 4:59-68.

94. Mutalik, V. K., P. Shah, and K. V. Venkatesh. 2003. Allosteric interactions and bifunctionality make the response of glutamine synthetase cascade system of Escherichia coli robust and ultrasensitive. *J Biol. Chem.*

95. Nacher, J. C., J. M. Schwartz, M. Kanehisa, and T. Akutsu. 2006. Identification of metabolic units induced by environmental signals. *Bioinformatics* 22:e375-e383.

96. Nielsen, J. 2001. Metabolic Engineering. *Applied Microbiology and Biotechnology* 55:263-283.

97. Nielsen, J. 2002. Metabolic engineering. *In* Encyclopedia of Physical Science and Technology. R. Meyers, editor. Academic Press, 391-406.

98. Nielsen, J. and S. Oliver. 2005. The next wave in metabolome analysis. *Trends Biotechnol.* 23:544-546.

99. Nielsen, J. and L. Olsson. 2002. An expanded role for microbial physiology in metabolic engineering and functional genomics: moving towards systems biology(1). *FEM. Yeast Res.* 2:175-181.

100. Nielsen, J. 2003. It Is All about Metabolic Fluxes. *J. Bacteriol.* 185:7031-7035.

101. Nissen, T. L., M. C. Kielland-Brandt, J. Nielsen, and J. Villadsen. 2000. Optimization of ethanol production in Saccharomyces cerevisiae by metabolic engineering of the ammonium assimilation. *Metab Eng* 2:69-77.

102. Nissen, T. L., U. Schulze, J. Nielsen, and J. Villadsen. 1997. Flux distributions in anaerobic, glucose-limited continuous cultures of Saccharomyces cerevisiae. *Microbiology* 143:203-218.

103. Overkamp, K. M., B. M. Bakker, P. Kotter, A. van Tuijl, S. de Vries, J. P. van Dijken, and J. T. Pronk. 2000. In Vivo Analysis of the Mechanisms for Oxidation of Cytosolic NADH by Saccharomyces cerevisiae Mitochondria. *J. Bacteriol.* 182:2823-2830.

104. Pal, C., B. Papp, and L. D. Hurst. 2001. Highly Expressed Genes in Yeast Evolve Slowly. *Genetics* 158:927-931.

105. Pal, C., B. Papp, and M. J. Lercher. 2006. An integrated view of protein evolution. *Nat Rev Genet* 7:337-348.

106. Palsson, B. O., N. D. Price, and J. A. Papin. 2003. Development of network-based pathway definitions: the need to analyze real metabolic networks. *Trends in Biotechnology* 21:195-198.

107. Papp, B., C. Pal, and L. D. Hurst. 2004. Metabolic network analysis of the causes and evolution of enzyme dispensability in yeast. *Nature* 429:661-664.

108. Patil, K. R., M. Akesson, and J. Nielsen. 2004. Use of genome-scale microbial models for metabolic engineering. *Current Opinion in Biotechnology* 15:64-69.

109. Patil, K. R. and J. Nielsen. 2005. Uncovering transcriptional regulation of metabolism by using metabolic network topology. *PNAS* 102:2685-2689.

110. Peltonen, L. and V. A. McKusick. 2001. GENOMICS AND MEDICINE: Dissecting Human Disease in the Postgenomic Era. *Science* 291:1224-1229.

111. Peregrin-Alvarez, J. M., S. Tsoka, and C. A. Ouzounis. 2003. The Phylogenetic Extent of Metabolic Enzymes and Pathways. *Genome Res.* 13:422-427.

112. Pfeiffer, T., I. Sanchez-Valdenebro, J. C. Nuno, F. Montero, and S. Schuster. 1999. METATOOL: for studying metabolic networks. *Bioinformatics.* 15:251-257.

113. Pharkya, P., A. P. Burgard, and C. D. Maranas. 2004. OptStrain: a computational framework for redesign of microbial production systems. *Genome Res.* 14:2367-2376.

114. Phelps, T. J., A. V. Palumbo, and A. S. Beliaev. 2002. Metabolomics and microarrays for improved understanding of phenotypic characteristics controlled by both genomics and environmental constraints. *Current Opinion in Biotechnology* 13:20-24.

115. Piper, M. D. W., P. Daran-Lapujade, C. Bro, B. Regenberg, S. Knudsen, J. Nielsen, and J. T. Pronk. 2002. Reproducibility of Oligonucleotide Microarray Transcriptome Analyses. An interlaboratory comparison using chemostat cultures of saccharomyces cerevisiae. *J. Biol. Chem.* 277:37001-37008.

116. Pramanik, J. and J. D. Keasling. 1997. Stoichiometric model of *Escherichia coli* metabolism: Incorporation of growth-rate dependent biomass composition and mechanistic energy requirements. *Biotechnol. Bioeng.* 56:398-421.

117. Prathumpai, W., J. B. Gabelgaard, P. Wanchanthuek, P. J. Van De Vondervoort, M. J. De Groot, M. McIntyre, and J. Nielsen. 2003. Metabolic control analysis of xylose catabolism in Aspergillus. *Biotechnol. Prog.* 19:1136-1141.

118. Price, N. D., J. L. Reed, and B. O. Palsson. 2004. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol.* 2:886-897.

119. Price, N. D., J. A. Papin, C. H. Schilling, and B. O. Palsson. 2003. Genome-scale microbial in silico models: the constraints-based approach. *Trends in Biotechnology* 21:162-169.

120. Raamsdonk, L. M., B. Teusink, D. Broadhurst, N. Zhang, A. Hayes, M. C. Walsh, J. A. Berden, K. M. Brindle, D. B. Kell, J. J. Rowland, H. V. Westerhoff, D. K. van, and S. G. Oliver. 2001. A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat Biotechnol.* 19:45-50.

121. Reed, J. L., T. D. Vo, C. H. Schilling, and B. O. Palsson. 2003. An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). *Genome Biol.* 4:R54.

122. Rossell, S., C. C. van der Weijden, A. L. Kruckeberg, B. M. Bakker, and H. V. Westerhoff. 2005. Hierarchical and metabolic regulation of glucose influx in starved Saccharomyces cerevisiae. *FEMS Yeast Res.* 5:611-619.

123. Rutter, J., M. Reick, L. C. Wu, and S. L. McKnight. 2001. Regulation of Clock and NPAS2 DNA Binding by the Redox State of NAD Cofactors. *Science* 293:510-514.

124. Samal, A., S. Singh, V. Giri, S. Krishna, N. Raghuram, and S. Jain. 2006. Low degree metabolites explain essential reactions and enhance modularity in biological networks. *BMC Bioinformatics* 7:118.

125. Sanford, K., P. Soucaille, G. Whited, and G. Chotani. 2002. Genomics to fluxomics and physiomics -- pathway engineering. *Current Opinion in Microbiology* 5:318-322.

126. Schilling, C. H., D. Letscher, and B. O. Palsson. 2000. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J Theor. Biol.* 203:229-248.

127. Schilling, C. H., M. W. Covert, I. Famili, G. M. Church, J. S. Edwards, and B. O. Palsson. 2002. Genome-Scale Metabolic Model of Helicobacter pylori 26695. *J. Bacteriol.* 184:4582-4593.

128. Schuster, S., S. Klamt, W. Weckwerth, F. Moldenhauer, and T. Pfeiffer. 2002a. Use of network analysis of metabolic systems in bioengineering. *Bioprocess Biosystems Engg* 24:363-372.

129. Schuster, S., D. A. Fell, and T. Dandekar. 2000. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat. Biotechnol.* 18:326-332.

130. Schuster, S., S. Klamt, W. Weckwerth, F. Moldenhauer, and T. Pfeiffer. 2002b. Use of network analysis of metabolic systems in bioengineering. *Bioprocess Biosystems Engg* 24:363-372.

131. Segal, E., M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.* 34:166-176.

132. Segre, D., D. Vitkup, and G. M. Church. 2002. Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. U. S. A* 99:15112-15117.

133. Sheikh, K., J. Forster, and L. K. Nielsen. 2005. Modeling hybridoma cell metabolism using a generic genome-scale metabolic model of Mus musculus. *Biotechnol. Prog.* 21:112-121.

134. Sherlock, G. 2000. Analysis of large-scale gene expression data. *Curr. Opin. Immunol.* 12:201-205.

135. Shlomi, T., O. Berkman, and E. Ruppin. 2005. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *PNAS* 102:7695-7700.

136. Stein, S. 1999. An integrated method for spectrum extraction and compound identification from GC/MS data. *J Am Soc Mass Spectrom* 10:770-781.

137. Stelling, J., S. Klamt, K. Bettenbrock, S. Schuster, and E. D. Gilles. 2002. Metabolic network structure determines key aspects of functionality and regulation. *Nature* 420:190-193.

138. Stephanopoulos, G., H. Alper, and J. Moxley. 2004. Exploiting biological complexity for strain improvement through systems biology. *Nat. Biotechnol.* 22:1261-1267.

139. Stephanopoulos, G., A. A. Aristidou, and J. Nielsen. 1998. Metabolic engineering Principles and methodologies. Academic Press, San Diego.

140. Stephanopoulos, G. 1999. Metabolic Fluxes and Metabolic Engineering. *Metabolic Engineering* 1:1-11.

141. Strogatz, S. H. 2001. Exploring complex networks. *Nature* 410:268-276.

142. Sturn, A., J. Quackenbush, and Z. Trajanoski. 2002. Genesis: cluster analysis of microarray data. *Bioinformatics* 18:207-208.

143. Sumner, L. W., P. Mendes, and R. A. Dixon. 2003. Plant metabolomics: large-scale phytochemistry in the functional genomics era. *Phytochemistry* 62:817-836.

144. Tai, S. L., V. M. Boer, P. ran-Lapujade, M. C. Walsh, J. H. de Winde, J. M. Daran, and J. T. Pronk. 2004. Two-dimensional transcriptome analysis in chemostat cultures: combinatorial effects of oxygen availability and macronutrient limitation in Saccharomyces cerevisiae. *J. Biol. Chem.*M410573200.

145. ter Kuile, B. H. and H. V. Westerhoff. 2001. Transcriptome meets metabolome: hierarchical and metabolic regulation of the glycolytic pathway. *FEBS Lett.* 500:169-171.

146. Tirosh, I., A. Weinberger, M. Carmi, and N. Barkai. 2006. A genetic signature of interspecies variations in gene expression. *Nat Genet* 38:830-834.

147. Tomita, M. 2001. Whole-cell simulation: a grand challenge of the 21st century. *Trends Biotechnol.* 19:205-210.

148. Townsend, J. P., D. Cavalieri, and D. L. Hartl. 2003. Population genetic variation in genome-wide gene expression. *Mol. Biol. Evol.* 20:955-963.

149. Tu, B. P., A. Kudlicki, M. Rowicka, and S. L. McKnight. 2005. Logic of the Yeast Metabolic Cycle: Temporal Compartmentalization of Cellular Processes. *Science* 310:1152-1158.

150. Vallino, J. J. and G. Stephanopoulos. 2000. Metabolic flux distributions in Corynebacterium glutamicum during growth and lysine overproduction. Reprinted from Biotechnology and Bioengineering, Vol. 41, Pp 633-646 (1993). *Biotechnol. Bioeng.* 67:872-885.

151. Van Winden, W. A., W. M. Van Gulik, D. Schipper, P. J. Verheijen, P. Krabben, J. L. Vinke, and J. J. Heijnen. 2003. Metabolic flux and metabolic network analysis of Penicillium chrysogenum using 2D [13C, 1H] COSY NMR measurements and cumulative bondomer simulation. *Biotechnol. Bioeng.* 83:75-92.

152. varma, A., B. W. Boesch, and B. O. Palsson. 1993. Stoichiometric interpretation of Escherichia coli glucose catabolism under various oxygenation rates. *Appl. Environ. Microbiol.* 59:2465-2473.

153. Varma, A. and B. O. Palsson. 1994. Metabolic Flux Balancing: Basic Concepts, Scientific and Practical Use. *Nat Biotech* 12:994-998.

154. Villas-Boas, S. G., M. Kesson, and J. Nielsen. 2005a. Biosynthesis of glyoxylate from glycine in Saccharomyces cerevisiae. *FEMS Yeast Res.* 5:703-709.

155. Villas-Boas, S. G., S. Mas, M. Akesson, J. Smedsgaard, and J. Nielsen. 2005b. Mass spectrometry in metabolome analysis. *Mass Spectrom. Rev* 24:613-646.

156. Villas-Boas, S. G., J. F. Moxley, M. Akesson, G. Stephanopoulos, and J. Nielsen. 2005c. High-throughput metabolic state analysis: the missing link in integrated functional genomics of yeasts. *Biochem. J.* 388:669-677.

157. Vitkup, D., P. Kharchenko, and A. Wagner. 2006. Influence of metabolic network structure and function on enzyme evolution. *Genome Biology* 7:R39.

158. Wall, D. P., A. E. Hirsh, H. B. Fraser, J. Kumm, G. Giaever, M. B. Eisen, and M. W. Feldman. 2005. Functional genomic analysis of the rates of protein evolution. *PNAS* 102:5483-5488.

159. Wang, Z. X., J. Zhuge, H. Fang, and B. A. Prior. 2001. Glycerol production by microbial fermentation: a review. *Biotechnol. Adv.* 19:201-223.

160. Weckwerth, W., M. E. Loureiro, K. Wenzel, and O. Fiehn. 2004. Differential metabolic networks unravel the effects of silent plant phenotypes. *Proc. Natl. Acad. Sci. U. S. A* 101:7809-7814.

161. Westergaard, S. L., C. Bro, L. Olsson, and J. Nielsen. 2004. Elucidation of the role of Grr1p in glucose sensing by Saccharomyces cerevisiae through genome-wide transcription analysis. *FEMS Yeast Res.* 5:193-204.

162. Westergaard, S. L., A. P. Oliveira, C. Bro, L. Olsson, and J. Nielsen. 2006. A systems biology approach to study glucose repression in the yeast *Saccharomyces cerevisiae*. *Biotechnology and Bioengineering*.

163. Wiechert, W. 2002. Modeling and simulation: tools for metabolic engineering. *Journal of Biotechnology* 94:37-63.

164. Zhang, Q., D. W. Piston, and R. H. Goodman. 2002. Regulation of Corepressor Function by Nuclear NADH. *Science* 295:1895-1897.